Artificial Intelligence in Public Discourse

Studies by Michael Alexandrovsky, Hanna Algedri, Micaela Barkmann, Tim Bax, Antonia Becker, Kyra Breidbach, Rabea Breininger, Christian Burmester, Eddie Charmichael, Alina Deuschle, Dana Dix, Mekselina Doganc, Kai Dönnebrink, Luisa Drescher, Kristof Engelhardt, Milan Ewert, Franziska Gellert, Kristin Gnadt, Nikolai Godt, Thiago Goldschmidt, Michelle Görlitz, Isabel Grauwelman, Yannick Hardt, Deborah Häuser, Jara Herwig, Malte Heyen, Till Holzapfel, Fabian Imkenberg, Virginia Jagusch, Ali Jandaghi, Joline Janz, Maximilian Kalcher, Karina Khokhlova, Paula Kirmis, Johanna Kopetsch, Ralf Krüger, Julia Laudon, Lina Lazik, Elen Le Foll, Sönke Lülf, Anna Ricarda Luther, Clara Matheis, Richard Matschke, Louisa Maubach, Sascha L. Mühlinghaus, Pia Münster, Felix Naujoks, Janeke Nemitz, Thimo Neugarth, Sarah Neuhoff, Till Nicke, Alina Ohnesorge, Lilith Okonnek, Cosima Oprotkowitz, Elisa Palme, Florian Pätzold, Tom Pieper, Daniel Pietschke, Ivan Polivanov, Muhammad Faraz Rajput, Janine Reichmann, Clara Schier, Rebekka Schlenker, Sabine Scholle, Archana Singh, Lennard Smyrka, Julia Stormborn, Konstantin Strömel, Johanna Tamm, Kim Targan, Hedye Tayebi, Franka Timm, Lisa Titz, Lea Tiyavorabun, Katharina Trant, Nikola Tsarigradski, Eva von Butler, Anita Wagner, Jasmin L. Walter, Nele Werner, Christoph Werries & Inga Wohlert, co-edited by Jacqueline Näther, Matthias Richter, Marcel Ruland & Anna Wiedenroth

Edited by

Tobias Thelen

PICS – Publications of the Institute of Cognitive Science
Universität Osnabrück

PICS – Publications of the Institute of Cognitive Science 2022, Number 03

Artificial Intelligence in Public Discourse

Studies by Michael Alexandrovsky, Hanna Algedri, Micaela Barkmann, Tim Bax, Antonia Becker, Kyra Breidbach, Rabea Breininger, Christian Burmester, Eddie Charmichael, Alina Deuschle, Dana Dix, Mekselina Doganc, Kai Dönnebrink, Luisa Drescher, Kristof Engelhardt, Milan Ewert, Franziska Gellert, Kristin Gnadt, Nikolai Godt, Thiago Goldschmidt, Michelle Görlitz, Isabel Grauwelman, Yannick Hardt, Deborah Häuser, Jara Herwig, Malte Heyen, Till Holzapfel, Fabian Imkenberg, Virginia Jagusch, Ali Jandaghi, Joline Janz, Maximilian Kalcher, Karina Khokhlova, Paula Kirmis, Johanna Kopetsch, Ralf Krüger, Julia Laudon, Lina Lazik, Elen Le Foll, Sönke Lülf, Anna Ricarda Luther, Clara Matheis, Richard Matschke, Louisa Maubach, Sascha L. Mühlinghaus, Pia Münster, Felix Naujoks, Janeke Nemitz, Thimo Neugarth, Sarah Neuhoff, Till Nicke, Alina Ohnesorge, Lilith Okonnek, Cosima Oprotkowitz, Elisa Palme, Florian Pätzold, Tom Pieper, Daniel Pietschke, Ivan Polivanov, Muhammad Faraz Rajput, Janine Reichmann, Clara Schier, Rebekka Schlenker, Sabine Scholle, Archana Singh, Lennard Smyrka, Julia Stormborn, Konstantin Strömel, Johanna Tamm, Kim Targan, Hedye Tayebi, Franka Timm, Lisa Titz, Lea Tiyavorabun, Katharina Trant, Nikola Tsarigradski, Eva von Butler, Anita Wagner, Jasmin L. Walter, Nele Werner, Christoph Werries & Inga Wohlert, co-edited by Jacqueline Näther, Matthias Richter, Marcel Ruland & Anna Wiedenroth

Edited by

Tobias Thelen

Tobias Thelen (ed.). 2022. *Artificial Intelligence in Public Discourse* (PICS – Publications of the Institute of Cognitive Science 2022, Number 03). Osnabrück: Osnabrück University, Institute for Cognitive Science.

This title can be downloaded at:
https://osnadocs.ub.uni-osnabrueck.de/handle/urn:nbn:de:gbv:700-202111125625

Layout and workflows based on awesome templates and tools by Language Science Press, https://langscipress.org

Cover and concept of design: Ulrike Harbort
Fonts: Libertinus, Arimo, DejaVu Sans Mono
Typesetting software: XƎLaTeX

Published by:
Institute of Cognitive Science
Osnabrück University
49069 Osnabrück
Germany
https://ikw.uni-osnabrueck.de

Storage and cataloguing done by Osnabrück University, Universitätbibliothek.

Contents

I The general public

II AI and subject experts

III The media

IV Politics, governments and non-government organizations

V Movies and literature

Vorwort

Dieses Buch ist im Rahmen des Cognitive-Science-Seminars “AI in public discourse” entstanden. Dass es zustandegekommen ist, ist dem außerordentlich großen Einsatz der Studierenden zu verdanken, die sich mit großem Engagement auf eine experimentelle Veranstaltungsform eingelassen haben.

Experimentell war das Seminar in mehrerlei Hinsicht. Zum einen fand es im Wintersemester 2020/2021 während der Corona-Pandemie statt und hatte daher mit verschiedenen Einschränkungen zu kämpfen. Grundsätzlich war das Seminar als “Hybrid-Seminar” angelegt, d.h. die Studierenden konnten wählen, ob sie an den wöchentlichen Sitzungen in Präsenz oder online teilnehmen. Die Möglichkeit von Präsenzsitzungen wurde aber im Laufe des Semesters wieder zurückgenommen, so dass der größte Teil als Online-Veranstaltung ablief.

Die zweite Besonderheit war die sehr große Teilnehmerzahl. Aufgrund anderer, entfallender Veranstaltungen und des für das Modul “Künstliche Intelligenz” ungewöhnlichen, weniger technischen Zuschnitts war das Interesse an dem Seminar sehr groß und insgesamt haben über 100 Studierende teilgenommen. Diskussionen im Plenum waren somit kaum möglich, stattdessen fand der Großteil der Arbeit in selbstorganisierten Kleingruppen statt, die sich über Online-Kommunikationskanäle ausgetauscht haben. Die regelmäßigen Termine wurden als “Walk-In-Sessions” mit einem kleingruppenorientierten Online-Videkonferenz-Tool (wonder.me) gestaltet, bei denen Studierende in ihren Gruppen arbeiten, Fragen stellen, sich gegenseitig Hilfestellung leisten oder sich einfach nur über beliebige Themen austauschen konnten. Dabei war es Ihnen freigestellt, ob Sie überhaupt teilnehmen, und ob sie zwei Stunden bleiben oder nur für eine kurze Frage hereinschauen.

Die dritte Besonderheit lag in der für die Studierenden ungewohnten methodischen Herangehensweise. Das Seminar hat ein grundsätzliches Thema vorgegeben, innerhalb dessen Kleingruppen ein eigenes Thema finden und dazu eine eigene explorative Studie durchführen und in einem wissenschaftlichen Paper dokumentieren sollten. Bei diesem Thema ging es darum, den öffentlichen Diskurs um das momentan sehr präsente Thema “Künstliche Intelligenz” aus einer fachlich informierten Perspektive zu beleuchten. Cognitive-Science-Studierende haben mindests grundlegende, häufig aber auch schon fortgeschrittene Kenntnisse aktueller KI-Technologien und -Methoden. Neu war für die meisten von ihnen die explorierende, meist qualitative Auswertung von nicht-wissenschaftlichen Texten, Videos und anderen Diskursbeiträgen *über* Künstliche Intelligenz.

Der letzte experimentelle Aspekt lag darin, dass im Seminar veröffentlichenbare Ergebnisse entstehen und die Teilnehmenden aus eigener Erfahrung typische Schreib-, Redaktions- und Publikationsprozesse kennenlernen sollten. Dies umfasste die Nutzung einer kollaborativen LaTeX-Umgebung (Overleaf über die Academic Cloud Niedersachsen), die Vorgabe einer LaTeX-Vorlage und eines Styleguides (LanguageSciencePress) sowie die Einhaltung von Deadlines und einen gegenseitigen Reviewing-Prozess. Ob die so entstandenen Kapitel dann tatsächlich veröffentlicht werden sollen, stand den Studierenden frei, die Entscheidung wurde individuell nach Bekanntgabe der Noten abgefragt. Auf Wunsch waren auch ein Pseudonym oder der Verzicht auf Nennung des eigenen Namens möglich.

Das Seminar umfasste die üblichen 14 Semesterwochen und war mit zwei Semesterwochenstunden und vier Leistungspunkten angesetzt. Es verlief in drei Phasen:

1. Methodische Grundlagen, Begriffsdefinitionen und Themenfindung
2. Durchführung der Studien und Textproduktion
3. Gegenseitiges Reviewing und Endredaktion

In der ersten Phase haben wir uns um ein gemeinsames Verständnis der Thematik bemüht, d.h. im Wesentlichen die Begriffe "AI" und "Public Discourse" abzustecken versucht. Die Ergebnisse dieser Phase sind in Kapitel 2 festgehalten. Im Verlaufe der engagierten Diskussionen ist deutlich geworden, dass es auch innerhalb des Studienganges Cognitive Science sehr unterschiedliche Sichtweisen auf den Themenkomplex "Künstliche Intelligenz" gibt, was einen guten Ausgangspunkt für die eigenen Studien in Phase zwei darstellte. Wir haben nicht "die" Definition von KI gesucht oder gar festgelegt, sondern festgestelt, dass es eine große Bandbreite von Verwendungsweisen gibt, die nur zum Teil mit den Ansätzen und Inhalten der wissenschaftlichen Disziplin "Künstliche Intelligenz" kongruent sind. Diese Feststellung ermöglicht interessante Perspektiven auf die Frage, ob es richtige bzw. falsche Verwendungen des Begriffes gibt und wer darüber entscheidet. Letztendlich war die Perspektve für das Seminar klar, dass wir jede Verwendung des Begriffs in öffentlichen Diskursen als untersuchenswert betrachten. Das Spannungsverhältnis zur wissenschaftlichen

Informatik-Teildisziplin “Künstliche Intelligenz” sollte dabei aber mitbetrachtet werden, da sich das Seminar als Lehrveranstaltung zu genau dieser Diszplin versteht.

Unter “Public Discourse” verstehen wir entsprechend eines solchen weiten Untersuchungsansatzes hier alle für die breitere Öffentlichkeit bestimmten Beiträge, d.h. keine rein privaten Äußerungen und keine wissenschaftlichen Fachpublikationen. Intensiv wurde die Frage diskutiert, wie viel Vereinheitlichung im Studiendesign und in der Studiendurchführung wünschenswert bzw. erreichbar sind. In der ersten Seminarphase haben wir an einer gemeinsamen Ontologie im Sinne eines hierarchischen Begriffsinventars gearbeitet, um auf das Ziel hinzuarbeiten, gut vergleichbare und aufeinander beziehbare Ergebnisse in Phase zwei erarbeiten zu können. Schnell hat sich allerdings gezeigt, dass dieses Ziel nicht realistisch war: Die Variationsbreite der inhaltlichen und methodischen Interessen der Teilnehmenden war so groß, dass alle Vereinheitlichungsversuche, z.B. die Beschränkung auf Diskursbeiträge in Textform oder Beiträge in englischer Sprache dazu geführt hätten, interessante und mit großem Engagement geplante Studien und Forschungsfragen schon vor Beginn ad acta zu legen. Wir haben uns daher entschlossen, die gemeinsam erarbeitete Ontologie eher als Orientierungshilfe denn als gemeinsame Struktur zu verstehen. Daher finden sich sowohl Beiträge, die sich ganz explizit auf diese Ontologie beziehen als auch Beiträge, die sich eher implizit davon haben inspirieren lassen.

Die Gruppen waren entsprechend sehr frei, sich eigene Forschungsfragen zu suchen. Ihnen wurden einige Beispiele für mögliche Fragen präsentiert, die grundsätzlich nach dem Muster aufgebaut waren: “Wie wird der Begriff ‘Künstliche Intelligenz’ in Bereich XY des öffentlichen Diskurses verwendet?”. Beispiele für solche Bereiche waren “Wissenformate im deutschen Kinderfernsehen” oder “Zeitungsreaktionen auf Alexa-Datenlecks”. Die Phantasie der Studierenden war hier allerdings bedeutend größer als die des Dozenten, so dass viele kreative und lohnende Ideen nur deshalb realisiert wurden, weil das Seminar letztlich thematisch sehr offen gestaltet war. Auch das methodische Herangehen an die Forschungsfragen blieb den Teilnehmenden überlassen. Genauer vorgesellt wurden qualitative Methoden aus dem Umfeld der qualitatischen Inhaltsanalyse, die zusammen mit möglichen Tagging-Tools diskutiert wurden. Eine Reihe von Gruppen haben aber quantitative Methoden bevorzugt oder sind methodisch etwas freier und explorativer vorgegangen.

Der Wechsel in Phase zwei brachte auch einen Wechsel des Seminarmodus mit sich: Es gab in dieser Phase nur wenige Online-Plenums-Sitzungen. Dabei haben alle Teams ihr Studiendesign und ihre Ergebnisse zweimal vorgestellt: Einmal zu Beginn der Phase und einmal gegen Ende. Außerhalb dieser Sitzungen haben die Studierenden selbstorganisiert in Kleingruppen gearbeitet und konnten und sollten individuelle Termin mit dem Dozenten vereinbaren und Online-Walk-In-Sessions nutzen, in denen sowohl Gruppenarbeit als auch Einzelgespräche möglich waren. Auf diese Weise konnten trotz der Seminarverhältnisse sprengenden Veranstaltungsgröße ein Überlick für alle und vor allem eine individuelle Betreuung in gewissen Maße gewährleistet werden. Die meisten Fragen und Anliegen bei den individuellen Gesprächen bezogen sich auf Fragen des methodischen Vorgehens. Viele Teilnehmende, insbesondere aus dem Bachelor-Studiengang, haben in diesem Seminar erstmalig an einer etwas größeren und eigenständig definierten Forschungsfrage gearbeitet, so dass erwartbare Unsicherheiten darüber bestanden, ob das in der Gruppe ausgehandelte Vorgehen korrekt und sinnvoll ist.

Phase zwei endete mit einer frühen Deadline, bis zu der die Studien durchgeführt und als ca. 10-seitiges Paper schriftlich festgehalten worden sein sollten. Da viele Gruppen sich sehr ambitionierte Themen und Methodiken vorgenommen haben, konnte der Termin trotz Verlängerung der Phase nicht durchgängig eingehalten werden. Die dritte Phase des gemeinschaftlichen gegenseitigen Reviews hat dann trotzdem bereits begonnen und alle Gruppen waren aufgefordert, deutlich kenntlich zu machen, wenn ihre Kapitel noch keinen finalen Stand erreicht hatten.

Die Reviewing-Phase muss als außerordentlich erfolgreich und konstruktiv bewertet werden. Auch hier gab es wenige Vorgaben für die Studierenden. Als Richtwert wurde ihnen lediglich eine zwei bis drei-stündige Beschäftigung mit den Texten anderer Gruppen aufgegeben, bei der sichtbare und möglichst hilfreiche Kommentare entstehen sollten. Es stand den Studierenden frei, ob sie einzelne Texte intensiver lesen und kommentieren, Einzelaspekte im Querschnitt über mehrere Gruppen betrachten oder ganz eigene Vorgehensweisen entwickeln wollen. Insgesamt haben die Studierenden deutlich über 3.000 Kommentare zu Texten ihrer Kommiliton:innen verfasst, die von Hinweisen auf Rechtschreibfehler über stilistische Verbesserungsvorschläge und Verständnisfragen bis hin zu sehr eingehenden Auseinandersetzungen mit methodischen und inhaltlichen Aspekten reichten. Damit sind sie im Durchschnitt sicherlich deutlich über die zeitliche Vorgabe von zwei bis drei Stunden hinausgegangen. Die Kommentare wurden direkt an den LaTeX-Quelltexten in der Overleaf-Umgebung vorgenommen und in vielen Fällen haben sich in dieser Kommentarspalte längere Diskussionen zwischen Reviewer:innen und Autor:innen ergeben. Bei der Auswertung des Kommentare, aber auch der Beratung der Gruppen ist durchgängig ein sehr konstruktives Verhalten der Studierenden aufgefallen, das deutlich darauf abzielte, anderen Gruppen bei der Verbesserung der

Texte ernsthaft helfen zu wollen. Konkurrenzgedanken haben dem Augenschein nach keine Rolle gespielt.

Vier Teilnehmende des Seminars haben keine eigenen Studien durchgeführt, sondern eine Editor:innen-Rolle eingenommen. In Phase zwei haben sie das LaTeX-Template für die Einzelkapitel vorbereitet und die Gruppen vor allem bei LaTeX-Fragen unterstützt und Korrekturen vorgenommen. In der Reviewing-Phase haben sie darauf geachtet, dass alle Texte mit Kommentaren versehen wurden und haben insbesondere auf formale Aspekte wie der korrekten Einbindung von Zitaten, LaTeX-Konventionen und anderen Aspekten des Style-Guides geachtet.

Dieser gemeinsamen Reviewingphase war auch ein frühes und festes Datum als Abgabezeitpunkt der Texte geschuldet: Bereits zwei Wochen nach Ende der Vorlesungszeit und nicht wie sonst bei Hausarbeiten üblich erst zum Ende der Semesterferien waren die Endfassungen fällig, weil nur so eine gemeinsame dritte Phase möglich war. Die Arbeitsintensität hat während des Seminars stetig zugenommen und hat bei nahezu allen Gruppen zum Ende hin das für ein Seminar diesen Umfangs übliche Maß überschritten.

Das Ergebnis dieses intensiven und von den Studierenden mit bewunderswertem Elan durchgeführten Seminars halten Sie nun in den Händen. Die enthaltenen Texte sind nicht perfekt, sie können und sollen es auch gar nicht sein: Es sind nicht mehr und nicht weniger als eigenständige Versuche der Studierenden, sich einer interdisziplinären Fragestellung zu widmen und dabei selbst forschend tätig zu werden. Alle in diesem Band versammelten Texte haben ihre ganz eigenen Stärken und stellen nicht nur als Ergebnisse von Lernprozessen, sondern ganz dezidiert auch aus Forschungsperspektive beachtenswerte Produkte dar. Sie haben selbstverständlich auch Schwächen und dürften in einigen Fällen methodisch rigoroser umgesetzt oder sprachlich ausgefeilter formuliert worden sein. Als Verantwortlicher für dieses Seminar kann ich sagen, dass diese Schwächen zum größten Teil auf die von mir gesetzten, stark einschränkenden Rahmenbedingungen der Veranstaltung zurückzuführen sind: Mit mehr Zeit und intensiverer Betreuung meinerseits wären die Ergebnisse auch noch um diese kleinen Mängel bereinigt worden.

Ich finde diese Schwächen insgesamt aber ganz unerheblich: Das vorliegende Buch ist aus meiner Sicht ein überwältigender Beweis für die Kreativität, den Forschungsdrang und die Begeisterung der Studierenden für aktives forschendes Lernen. Es enthält hunderte von Ideen, Argumenten, Ergebnissen und Sichtweisen, die mich als Dozenten und Forscher überrascht und begeistert haben.

Die enthaltenen Kapitel sind keine “Auswahl der Besten”, sondern enthalten alle Abgaben, deren Veröffentlichung alle Beteiligten nach Bekanntgabe der Noten freiwillig zugestimmt haben und die den von mir gesetzten qualitativen Standards genügen. Somit zeigen sie die ganze Bandbreite an Themen, Herangehensweisen und auch Suchprozessen aus dem Seminar. Als Herausgeber dieses Bandes stehe ich hinter jedem einzelnen Beitrag und halte noch einmal fest: Das, was die Beiträge leisten, verdanken sie einzig und allein dem Engagement der Autor:innen und Reviewer:innen, was an ihnen unperfekt und lückenhaft ist, ist meiner kaum zureichenden Betreuung geschuldet.

Vielen Dank an alle, die zu diesem Buch beigetragen haben, es war mir eine große Freude, mit Euch zusammenarbeiten zu dürfen!

Ein besonderer Dank geht an die vier Editor:innen:

- Jacqueline Näther
- Matthias Richter
- Marcel Ruland
- Anna Wiedenroth

Nicht zuletzt danke ich dem Verein der Freunde und Förderer des Instituts für Kognitionswissenschaft e.V., der es möglich gemacht hat, nicht nur eine PDF-Fassung zu produzieren, sondern diese Sammlung auch als echtes Buch zu drucken. Es macht einen großen Unterschied, die Ergebnisse der eigenen Arbeit auch tatsächlich in den Händen halten zu können.

Tobias Thelen, Osnabrück im Juni 2022

Preface

This book was produced as a result of the Cognitive Science seminar “AI in public discourse”. The fact that it came to life is due to the extraordinary commitment of the students, who got involved in a very experimental form of teaching and learning with great commitment.

The seminar was experimental in several respects. First, it took place in the 2020/2021 winter semester during the Corona pandemic and therefore had to overcome various limitations. Originally, the seminar was designed as a “hybrid seminar”, i.e., students could choose whether to attend the weekly sessions in face-to-face or online. However, the option of face-to-face attendance was withdrawn as the semester progressed, so that most of the course was conducted online.

The second issue was the very large number of participants. Due to other courses being cancelled and because this course was an unusual and less technical one for the module “Artificial Intelligence”, the interest in the seminar was very high and more than 100 students attended. Discussions in the plenum were therefore hardly possible, and instead, most of the work took place in self-organized small groups exchanging information via online communication channels. Most of the regular meetings were held as “walk-in sessions” with a small-group oriented online video tool (wonder.me), where students worked in their groups, asked questions, provided assistance to each other, or simply exchanged ideas on any topic. Students were always free to participate in these walk-in-sessions at all, and could stay for two hours or just drop in for a quick question.

The third special feature was the methodological approach, which was unfamiliar to most of the students. The seminar was conceived around an overarching topic, within which small groups find their own topic, conduct their own exploratory study, and document it in an academic paper. This overarching topic was about exploring the currently very topical subject of “Artificial Intelligence” in public discourse from a professionally informed perspective. Cognitive science students have at least basic, but often already quite advanced knowledge of current AI technologies and methods. New to most of them was the exploratory, mostly qualitative evaluation of non-scientific texts, videos, and other contributions to public discorse *about* AI.

The final experimental aspect lay in the fact that publishable results were expected to emerge from the seminar and the participants were to become acquainted with typical writing, editing and publishing processes first hand. This included the use of a collaborative LaTeXenvironment (Overleaf/ShareLaTeX via Academic Cloud Niedersachsen), the use of a LaTeXtemplate and style guide (Language Science Press) as well as the adherence to deadlines and a mutual reviewing process. The students were free to decide whether the resulting chapters should actually be published. The decision was made individually after the grades were announced. If desired, it was possible to use a pseudonym, or not to include a name at all.

The seminar comprised the usual 14 weeks within a semester and was scheduled as two semester hours and four credit points. It proceeded in three phases:

1. Methodological basics, definitions of terms, and identification of topics.
2. Conducting the studies and text production.
3. Mutual reviewing and final editing.

In the first phase, we sought to establish a common understanding of the subject matter, i.e., essentially to delineate the terms “AI” and “public discourse”. The results of this phase are to be found in chapter 2. In the course of engaged discussions it has become clear that even within the Cognitive Science program there are very different views on what constitutes “Artificial Intelligence”. This proved to be a good starting point for the own studies in phase two. We did not look for “the” definition of AI, but found that there is a wide range of uses, only some of which are consistent with the approaches and research topics of the scientific discipline of “Artificial Intelligence”. This observation provides interesting perspectives as to whether there are correct or incorrect uses of the term and who decides about that. Ultimately, the perspective for the seminar was clear; we considered any use of the term in public discourse to be worthy of investigation. The tension between the term and the scientific discipline of “Artificial Intelligence” was to be taken into account then, since the seminar was intended as a course on precisely this discipline.

By “public discourse” we mean, in accordance with such a broad approach to the research questions, all contributions intended for the public, i.e., not purely private statements and not scientific publications. The question was: how much standardization around study design and methods is desirable and how much is achievable. In the first seminar phase, we worked on a common ontology in the sense of a hierarchical conceptual inventory, in order to work towards the goal of achieving interrelated results in phase two. However, it quickly became apparent that this goal was not realistic. The variation of participants' content-related and methodological interests was so large that any attempts at standardization, e.g., limiting discourse contributions to text form or contributions in English, would have led to the exclusion of interesting studies and research questions, which had been planned with great engagement. We therefore decided to use the jointly developed ontology more as an orientation guide rather than a common structure. Therefore, there are chapters that explicitly refer to this ontology as well as chapters that were inspired by it rather implicitly.

Accordingly, the groups were very free to seek their own research questions. We presented and discussed some examples of possible questions in the beginning, basically along the lines of "How is the term 'artificial intelligence' used in domain XY of public discourse?". Examples of such areas were "Education formats in German children's television" or "Newspaper reactions to Alexa data leaks". Students' imagination was significantly greater than the lecturer's, however, so that many creative and worthwhile ideas were ultimatively only realized because the seminar was so open thematically. The methodological approaches were also left to the participants. Qualitative methods from the field of qualitative content analysis were presented in detail together with possible tagging tools in the seminar. However, a number of groups preferred quantitative methods or took a somewhat freer and more exploratory methodological approach.

The change to phase two also entailed a change in seminar mode: There were only four online plenary sessions, in which all teams presented their study design and results twice: Once at the beginning of the phase and once towards the end. Outside of these sessions, students worked in a self-organized manner in small groups and were encouraged to make individual appointments with the instructor and use the aforementioned walk-in sessions. In this way, it was possible in spite of the given size of the seminar to ensure an overview of all the studies, and above all, individual support could be guaranteed to a certain extent. Most of the questions and concerns raised during the individual discussions were related to methodological approaches. Many participants, especially from the bachelor's program, worked on a somewhat larger and independently defined research question for the first time in this seminar, so as expected, there were uncertainties about whether the approach negotiated in the group was correct and appropriate.

Phase two ended with an early deadline by which the studies had to be finished and written up as a 10-page paper. Since many groups had very ambitious topics and methodologies, the deadline could not be consistently met despite an extension of the phase. Nonetheless, the third phase of the collaborative peer review had to be started early and all groups were asked to clearly indicate if their chapters had not yet reached a final state.

The reviewing phase was extraordinarily successful and constructive. Here, too, there were few guidelines for the students. For orientation, they were told to spend two to three hours working on the texts of other groups, during which visible and helpful comments were to be produced. The students were free to decide whether they read and commented on a few texts more intensively, or looked at single aspects across many papers, or developed their own approaches. All in all, the students wrote well over 3,000 comments on the texts ranging from the correction of spelling mistakes, suggestions for stylistic improvements, and questions of comprehension to very detailed discussions of methodological and content-related aspects. Thus, on average, they have certainly gone far beyond the time limit of two to three hours. The comments were made directly on the LaTeXsource texts in the Overleaf environment and in many cases longer discussions between reviewers and authors took place in this comment column. From evaluating comments, a very constructive behavior of the students became visible, which was clearly aimed at seriously helping other groups to improve the texts. To all appearances, competitive thoughts did not play a counter-productive role.

Four seminar participants did not conduct their own studies, but took on an editor role. In phase two they prepared the LaTeXtemplate for the individual chapters and supported the groups mainly with LaTeXquestions and provided technical assistance. In the reviewing phase, they made sure that all texts were annotated and paid particular attention to formal aspects such as the correct integration of citations, LaTeXconventions and other aspects of the style guide.

This joint reviewing phase also had an early and fixed date as the deadline for the submission of the texts: The final versions were due as early as two weeks after the end of the lecture period and not at the end of the semester break, as is usually the case with term papers. The intensity of the work increased steadily during the seminar and exceeded the usual level for a seminar of this size for almost all groups towards the end.

The result of this intensive seminar, which was carried out by the students with admirable verve, is now in your hands. Some texts are not perfect, nor can they be, nor are they intended to be. They are no less than autonomous attempts by the students to address an interdisciplinary questions and to become active in their own research.

All texts collected in this volume have their own unique strengths and represent not only the results of learning processes, but are also valuable from a research perspective. Of course, they also have weaknesses and in some cases they could have been methodologically more rigorous or make more sophisticated use of scientific language. As the person responsible for this seminar, I can say that these weaknesses have to be attributed to the severely restrictive framework conditions of the event set by me: With more time and more intensive supervision, the results would have been corrected for these minor deficiencies as well.

However, I find these weaknesses insignificant: This book is, in my view, an overwhelming account of students' creativity, interest in research, and enthusiasm for active inquiry-based learning. It contains hundreds of ideas, arguments, results, and viewpoints that surprised and excited me.

The chapters included in this book are not a "selection of the best," but include all the submissions that the participants have voluntarily agreed to be published after the grades were announced, and which meet minimum quality standards. Thus, they show the full range of topics, approaches, and also search processes. As editor of this volume, I stand behind each and every contribution. The studies' achievements are solely due to the commitment of the authors. What is imperfect and incomplete in the contributions is due to my barely sufficient supervision.

A huge thank thanks to all who contributed to this book, it was a great pleasure to work with you!

Special thanks go to the four editors:

- Jacqueline Näther
- Matthias Richter
- Marcel Ruland
- Anna Wiedenroth

Last but not least, I would like to thank the Association of Friends and Sponsors of the Institute of Cognitive Science (F^{2}IKW e.V.), which made it possible not only to produce a PDF version, but also to print this collection as a real book. It makes a big difference to actually be able to hold the results of one's own work in one's hands.

Tobias Thelen, Osnabrück in June 2022

Introduction

Chapter 1

Introduction

Tobias Thelen

1 What is this book about?

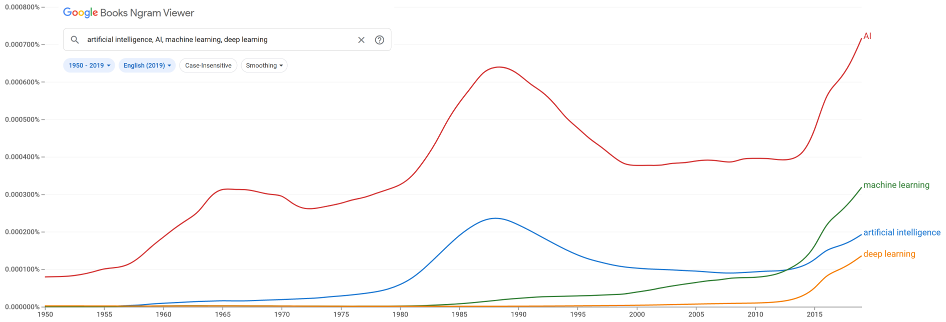


Figure 1: Frequency of AI terms used in books (1950-2019, Google Books ”English 2019” corpus). The terms are (from top to bottom): AI, machine learning, artificial intelligence and deep learning.

“Artificial Intelligence”, “AI” and other terms formerly were mere technical terms or used echoes of science fiction to describe distant, utopian states. Figure 1 shows the frequency of these terms in the Google Books corpus between 1950 and 2019. After a peak in the second half of the 1980s, it became clear that the promises of AI technology would not become reality, or at least not quickly. This changed in the mid-2010s, coinciding with the rise of the term “Machine Learning.” With the availability of really large amounts of data and new technical possibilities for efficient processing and analysis, numerous practical applications have also become visible to the public under the label ”Artificial Intelligence”.

Today, “Artificial Intelligence” is on everyone’s lips. The use of the term obviously differs significantly from its use in the computer science sub-discipline of the same name, which has been established since the late 1950s. In general, both narrower and broader uses of the term can be observed. Narrower in the sense that public uses of the term are mostly limited to machine learning and data-driven methods, but also include methods that would not be classified as AI in computer science. In addition, there are uses that are largely decoupled from the scientific discourse. On the one hand, AI is used as a marketing term that gives hardware and software an aura of superior technology but also of outstanding, slightly mysterious performance.

On the other hand, many AI terms are humanizing and evoke associations and equations with genuinely human performances. Terms such as “intelligence”, “learning”, or “recommendations” are used in everyday language to refer to people. If they are used unreflectively for technical artifacts, it is very easy to ascribe other human characteristics to them as well. Discussions about whether and when artificial intelligent machines will outperform or even dominate humans are an obvious consequence.

With the contributions in this book we want to examine the diversity of uses of the term “Artificial Intelligence” in public discourse and discuss it primarily against the background of the computer science discourse. We have divided the studies into four sections reflecting the agents of discours: The general public, the media, i.e. professional journalism, AI and subject experts and artists, especially those creating literature and movies.

2 The general public

The very first chapter already puts our conception of “public discourse” to the test. In their paper *“How do non-experts discuss Artificial Intelligence?”* (Chapter 3), Michael Alexandrovsky, Mekslina Doganc, Virginia Jagusch, Julias Stormborn, and Daniel Pietschke did not analyze existing public statements, but rather sat several groups of young, non-expert people around a table and had them discuss AI topics. It turned out that AI for these groups is hardly known as a sharply defined term, but is mixed up with other, even broader issues of digitalization and viewed with rather diffuse feelings of unease. This unease concerns both the presumed consequences of the further use of AI and their own knowledge of the subject, which they considered to be too low.

Many chapters in this book investigate newspaper articles or online-articles from newspaper sites. So does Sascha L. Mühlinghaus in his study *"What impact have experts on the public discourse about Artificial Intelligence?"*, (Chapter 4) but he does not primarily look at the articles themselves but at the public comments from readers. For an qualitative in-depth analysis of comments, he picked two articles from or with AI experts with a philosophical background. Mühlinghaus shows that it seems to be possible to evoke either more controversial or more in-depth discussions by choosing a certain style of presenting expert arguments. Especially for more controversial discussions, more misconceptions or overgeneralizations of the term "AI" occur.

Trying to understand general public discourse not only calls for detailed qualitative studies but also for using very large amounts of data to find out about general tendencies or to identify interesting discussions in the huge communicative spaces of social media. In their chapter *"Discourse about AI on Reddit"* (Chapter 5), Richard Matschke and Till Nicke explore quantitative methods for analyzing the billions of subreddits, posts and comments on Reddit. They developed Python tools to access and search the data, and extract subreddits presumably dealing with AI topics. It becomes apparent that the amounts of data at hand when using this kind of corpus cannot be fully cleaned or controlled but have to be interpreted and filtered carefully afterwards. For some very active channels, automatically generated word clouds are presented as a tool for an intuitive understanding of tendencies in large public discourse spaces.

Twitter is a good source for public discourses involving many different stakeholders. In their paper *"Twitter users' perception of artificial intelligence in remote proctoring"* (Chapter 6), Elen Le Foll and Lisa Titz analyzed a discourse that abruptly came into focus during the Corona pandemic. They automatically identified relevant tweets about online proctoring systems and incidents, manually tagged them in a detailed way, and analyzed them. They were able to show that Twitter is an important channel for stakeholders who may otherwise be barely heard. Students expressed concerns about the technological background of proctoring systems, but these were not addressed clearly enough by manufacturers and educational institutions. In particular, the actual role of AI technology in online proctoring was not made transparent according to the authors.

Twitter dicsussions were also investigated by F.K. and F.P. for their chapter *"How do Twitter users receive AI-related field research conducted by police?"* (Chapter 7). In this study, Tweets adressing the public test of face recognition technology by the Berlin police were analyzed using qualitative methods. The authors could show that mostly negative comments came up on Twitter, focussing on criticism on technical aspects like saftey and reliabilty of the systems.

3 AI and subject experts

The second part contains studies that analyze what experts say about AI adressing a broader public. This includes both AI experts as well as experts in a field that is presumably affected by AI advances. A highy influential medium distributing talks by experts with very different backgrounds are the TED talks. Antonia Becker, Michaelle Görlitz, Yannick Hardt and Clara Schier identified popular talks about AI and analysed them in their study *"How do scientists explain AI to the broad public? – A TED talks analysis"* (Chapter 8). By using quantitative and qualitative methods, they could show that in the TED talks they analyzed, experts with a STEM background tend to portray AI more positively while talks with a humanities background rather focus negative consequences.

Maximlian Kalcher takes a closer look at podcasts as another medium that gives a wide range of experts an opportunity to address the public. In his paper *"How is AI perceived amongst experts of different disciplines"* (Chapter 9), he investigates eight episodes of the popular Lex Fridman podcast and shows by applying AI speech recognition and key phrase identification how terms are more frequently used by either more optimistic or more pessimistic experts.

While the first two chapters in this part covered different experts appearing in a single medium or channel, Alina Deuschle and Joline Janz analyse appearances of a single expert in different media. The paper *"How does Elon Musk portray the dangers and the future of AI?"* (Chapter 10) analyses 17 longer popular videos containing interviews or discussion with Elon Musk. The authors show how Musk uses his popularity to talk about visions and outlooks on the future that sometimes are more opinions than facts backed by scientific findings.

Concluding this part, Janine Reichmann and Ali Jandaghi report about a qualitative analysis of opinions expressed in YouTube videos under the title *"How is AI in healthcare perceived by physicians in 2020?"* (Chapter 11). They searched for physicians expressing their opinions in different YouTube channels and formats. By analysing 20 videos, they found that mostly positive expectations about AI as a diagnotic and assisting tool were mentioned and neither fear of job losses nor worries about weaknesses of AI played an important role.

4 The media

The largest group of studies in the seminar focussed on analyzing discourse contributions by professional journalists in well-established media like newspapers or public television. This might be because these types of contributions are more easily retrievable, have a more homogenous form, and might thus be easier to analyze. But of course, professional media undoubtly do play an important role in public discourse as they have a high reach and are traditionally considered to be relevant players in public discourse.

Nikolai Godt, Ivan Polivanov, Karina Khokhlova and Mohammad Faraz Rajput open this part with a very interesting twist on the question at hand: They investigate how AI is perceived as a game-changer in journalism by being able to generate and personalize content. Their sutdy *"How AI transforms public discourse — An analysis of the impact of AI in public discourse as portrayed in major news outlets"* (Chapter 12) finds that these changes are closely watched and vividly discussed by traditional newspapers and that negative aspects like fake news and deepfakes play a big role in that discussion.

For Chapter 13, two groups teamed up during the semester. Sabine Scholle, Konstantin Strömel, Archana Singh, Louisa Maubach, Johanna Kopetsch, Kyra Breidbach, Kristin Gnadt, Anna Ricarda Luther, Lea Tiyavorabun and Hedye Tayebi looked at German, US american and Chinese newspapers to find about their title question: *"Does AI in public discourse change with different political and socio-economic systems?"*. They present evidence from severeal hundred articles they analyzed. Chinese newspapers in their analysis are the most positive in its discourse around AI, while the US and German media try communicate information on artificial intelligence mainly in an informative manner. The authors connect cultural and political tendencies of the respective countries with discourse aspects about strengths, weaknesses, opportunities and threats of AI.

Newspapers and news websites are mostly driven by notable events that may spark more in-depth discussion. A series of such events occured in connection with security and privacy incidents concerning voice assistents, namely the Amazon Alexa devices and services. Kim Targan, Clara Matheis and Kristof Engelhardt took a closer look at popular British news websites in their paper *"Ethical concerns and AI: Analysing British news articles about Alexa"* (Chapter 14). They show that ethical aspects are covered in the articles to a greater extend but some important questions are found to be missing because of a too limited technical understanding of the AI methods involved.

While news articles often are reserved about explaining technical details, Franziska Gellert, Julia Laudon, Pia Münster and Rebekka Schlenker investigated media that strive for the opposite. Their paper *"How is AI explained to children? A qualitative analysis of educational videos for children"* (Chapter 15) discusses seven videos from German public television adressing children. They found that the videos explain AI methods and technolgies quite accurate and use child-friendly examples but refrain from critial discussions and heavily humanize AI, especially robots.

Micaela Barkmann, Dana Dix and Kai Dönnebrink ask *"How is AI portrayed in Netflix' documentary 'The Social Dilemma' and how do newspapers react to it?"* (Chapter 16). By using extensive qualitative content analyses, they show how a very popular documentary was able to encourage a wide range of public discourse contributions. The mostly negative tone of the documentary was not simply picked up by journalists reporting about it but was reflected with additional arguments. This led to surprisingly diverse, differentiated and regionally diverging discussions.

The topic of AI in medical diagnostics was already covered in Chapter 11 where medical experts' opinions were analysed. Isabel Grauwelman, Cosima Oprotkowitz and Katharina Trant instead looked at news articles in their study *"A modern god complex - Doctor Who? An analysis of (un-)specialized German news articles on AI in medical diagnostics"* (Chapter 17) and analyzed sentiments, expressions of strengths and weaknesses and demands of actions. In line with the other group's findings, the authors were able to identify a mostly positive attitude towards the use of AI for diagnostics, both in spezialized and unspezialized articles. Threats and weaknesses were named but seem to be percevied as less important than the strengths and opportunities.

In 2020, a Nature article on an AI system for breast cancer screening sparked headlines like "Google's AI beats doctors at detecting breast cancer" all over the world. As these kinds of headlines tend to oversimplify scientific finding, Tim Bax, Milan Ewert, Florian Pätzold and Franka Timm wanted to find out more about how accurate the scientific study was received and reported in various media. Their study *"AI in healthcare – Expectation vs. reality of breast cancer detection"* (Chapter 18) identified a more critical discussion in blog articles than on news sites, but all in all, the authors found the overall public reception and disucssion in all articles to be balanced and accurate.

Apart from medical diagnostics, the use of robots in care is an often discussed topic in the public discourse on AI in healthcare. Rabea Breininger, Luisa Drescher, Lilith Okonnek and Inga Wohlert named their study *"Should robots take care of the elderly? – Comparing ethical guidelines to real life experiences* and conducted qualitative analyses using a SWOT tagset on official guidelines by the German Ethics Council and on twelve TV documentaries about experiences with robots in elderly care. The authors found many documentaries overgeneralizing the term "AI" by using it synonymously also for robot-like puppets in elderly care or digitalization in general. They found that much of the fears and negative attitudes expressed in the documentaries address technical features that are far beyond the current state of the art in AI.

But what happens if the unthinkable happens and AI systems prove to be able to do things assumed to be impossible before? This situation is addressed by Sarah Neuhoff, Ralf Krüger and Nikola Tsarigradski in their study *"Asian reaction to AI supremacy in the game of Go"*. Even after chess champions having no chances against advanced chess playing computers, the game of Go was considered to be too difficult and to require true human ingenuity. From a qualitative analysis of east asian news articles, the authors find suprise and the fear of job loss for professional Go player as the most common emotional reactions. But the event is also seen as an opportunity to explore new ways of playing Go and as a justification of intensifying AI research now that it has proven its power.

5 Politics, governments and non-gonvernemt organizations

The fourth part of the book examines how governments, political and non-government organizations communicate about AI.

Christian Burmester, Thiago Goldschmidt, Felix Naujoks and Tom Pieper start with their paper *"AI made in Germany"*, (Chapter 21) in which they analyze the AI strategy paper of the German government with mixed-method techniques. They find a high technical level and a broad awareness of opportunities and risks, but criticize that action plans might be too passive plans and that some arguments such as the presumed lack of acceptance in the population are not supported by scientific evidence.

Jara Herwig, Lina Lazik, Sönke Lülf and Elisa Palme extend this view to the international stage. In their study *"A comparison of different governmental approaches to prepare the public for the age of AI"* (Chapter 22), they examine government communication strategies and strategy papers from the USA, Japan, Australia and Finland. They find that in all four cases AI is primarily seen as a positive, important and promising field of development that should be promoted and that is also capable of solving specific problems of the respective country.

In addition to AI strategies of individual countries, there are also efforts to find internationally agreed regulations. In their study *"Ethical guidelines in the European judicial system"* (Chapter 23), Hanna Algedri and Till Holzapfel analyze a recommendation paper of the Council of Europe, a rather loose association of 47 European countries. It contains recommendations for AI technologies in judicial systems that should either be be expanded, carefully weighed, or used only with extreme caution. The authors elaborate that the charter under review presents AI technologies in an accurate way and comprehensively identifies the ethical implications.

Eva von Butler, Janeke Nemitz and Nele Werner compared two different views and intentions on the same topic in their paper *"AI as part of the energy transition – A comparison of the portrayal of AI from the big energy group E.ON and the non-profit organisation Germanwatch"* (Chapter 24). In a qualitative analysis of papers by both stakeholders, the authors find that both the company and the NGO identify chances in using AI for the energy transition and for fighting climate change. Risks, however are mainly stressed by the NGO and seem to be neglected by E.ON.

A closer look at the inner workings of politics take Eddie Charmichael and M.S. in their study *"How AI in the form of content filters for social media is discussed in the German parliament"* (Chapter 25). By mining the German parliament's meeting minutes, they found 44 utterances concerning upload filters which they analyszed with quantitative and qualitative methods. The authors show that the discussion did not manage to reach a productive level of proposing solutions. Instead, alarmistic exaggeration of risks was much more common in the parliament discussion that was driven by emotions and aims of influencing opinions.

6 Movies and literature

The book's last part covers art, namely movies and literature as a place of public discourse. Fabian Imkenberg, Paula Kirmis, Johanna Tamm and Christoph Werries take a general look at the *"Portrayal of AI in popular movies"* (Chapter 26). They selected four more recent movies from a list of highest-grossing science fiction movies (Her, Wall-E, Gost in the shell and I, Robot) and analyzed them with qualitative methods according to eight guiding questions of AI portrayal. The authors find AI depicted as a potential helper for humans in all the movies which on the other hand also raise the question of what differentiates human and artificial characters and how humanity as such could be defined.

A more specialized question concerning AI in movies taken from the same period is addressed by Thimo Neugarth, Alina Ohnesorge, Lennard Smyrka and Jasmin L. Walter in their paper *"AI Fatale – An analysis of AI characters focused on gender depiction and inflicted harm in movies from 2000–2020"* (Chapter 27). For gathering data from the top 15 highest-grossing movies and subsequently 21 AI characters, the authors apply a graded assessment of the AI representation (e.g. physical form, communication), gender depiction (e.g. physical form, voice impression and pronouns), harm inflicted by the AI (form of harming, motivation to harm), and the general movie setting and power dynamics separately for every movie third. They find a rather balanced distribution of gender in terms of harming behavior, but also uncovered problematic dynamics in gender depiction, specifically in respect to the motivation and form of harm inflicted by female AI characters, which seem to match typical stereotypes of female violence.

"AI in literature: An investigation of the science fiction novel QualityLand" is the title of the study by Malte Heyen, F.R. and Anita Wagner that concludes the book. They performed a mixed-method analysis of the German satiric and dystopian science fiction novel "QualityLand" and show how the novel traces some of the most prominent philosophical discussions on AI from past decades. The authors also find the novel to explain technical aspects in a precise, yet approachable way, and to depict AI in a quite versatile manner, as interactions and conversations with AI characters serve different purposes in the novel.

7 Overall conclusion. Or: What did we find out?

Given the extraordinarily diverse and multifaceted approaches of the studies presented in this book, it is difficult to draw a clear and pointed conclusion about AI in public discourse. Instead, only a few more general findings and tendencies will be mentioned, and, just like the preceding overview, should encourage the reader to venture a detailed look into the individual chapters.

Public discourse about AI is indeed taking place. The authors of the studies hardly had any problems finding relevant material for their analyses. AI is a topic that has evidently made it out of specialized scientific discussion and appeared prominently in the public in recent years.

Public discourse is relatively well informed technically. Gross mistakes, inadmissible simplifications or broad generalizations were found in various places, but not as a general tendency. Nevertheless, the AI concepts of the nearly 70-year-old computer science sub-discipline "artificial intelligence" and current non-scientific discourse are not congruent. Public discourse essentially equates AI with the application of machine learning techniques, and does not differentiate sharply between AI and more general data analysis techniques or the terms "data science" or "big data". In discourses distant from technology, any kind of algorithmic decision making or any kind of robotics is quite often referred to as AI.

It has been found that AI topics are treated in a differentiated manner in most cases. In almost all fields of discourse and contributions, both the presentation of strengths and opportunities as well as of weaknesses and risks could be found.

This overall quite balanced presentation found in the studies of this book could be a good basis for a productive public discourse. However, the prerequisite for this is a profound preoccupation with technical details, ethical questions and a differentiated consideration of application fields. The question is whether the public is willing and able to acquire these prerequisites and to create educational opportunities that are actually used and effective. This is a question for another book as it will be the subject of another collection of studies to be conducted in a seminar in the winter semester 2022/2023 entitled “Learning about AI”.

Chapter 2

The common ontology used in this book

Deborah Häuser & Archana Singh

In the following chapter the ontology developed for the purpose of this book will be covered. Many of the individual chapters applied this ontology as the guideline of their research. Before the book-specific ontology will be discussed further, the general role of ontologies in qualitative reasearch is explained through the example of codebooks. Lastly, the chapter will cover the development of the hierarchical sets of terms and concepts that have been discussed in the seminar.

Keywords: Ontology | Coding | Codebook

1 What is an ontology for qualitative research?

Qualitative research is a broad term describing a research approach that stands in opposition to quantitative research. It targets unstandardized, mainly unstructured data, for instance, interview protocols. Thereby, qualitative research is used to analyse social phenomena such as people's perspectives, concepts and opinions. A common method to analyse qualitative data is the process of coding based on an underlying codebook. The meaning of coding, as well as its aim, will be outlined in this section and afterwards compared to the properties of the ontology which was developed for this book. (Lapan et al. 2012: 42,43), Leavy (2014: ch. 1)

A codebook is a set of codes developed and used as a guide to analyse data. A code in turn is mostly a word or a sentence that assigns a label to a certain part of data. In the codebook the codes are explained by providing definitions and examples. Also, the relations between different codes can be illustrated. It is distinguished between "priori codes" which are codes that were developed before the study of the data began and “inductive code” which comprises codes that were developed during the research.

Through labelling specific parts, the researchers assign a certain interpretation to the labelled data. Thereby, the code represents and captures an interpretation. When working with other researchers, using a codebook increases consistency across a research field since it helps to identify and exemplifying data with the same underlying conceptual idea, i.e. the same interpretation. Therefore,codebooks simplify collaborative research as well as building upon already existing research.

The description of the codes as well as the codebook in general increase transparency since they make it easier for other people to retrace the research. Furthermore, the explanations of the codes explicate interpretations of the researcher and helps to reduce biases and implicit assumptions that can otherwise reduce the quality of the research. Labeling data with codes enables researchers to organize their raw data into more meaningful segments that can be further analyzed. Therefore, labeling the data reduces its complexity, and brings it into a format that is easier to further work on. Altogether, the coding process helps in determining whether the data support a specific research theory or not. Gibbs (see 2018: ch. 4), (Vaismoradi & Snelgrove 2019: see)

2 Our common ontology - What do we need an ontology for?

The ontology that is developed for this book shares some important features with the described codebooks. It provides words and phrases that are used to label the data. Thereby, it also aims to identify parts of the data that capture the same conceptual meaning as represented by the particular term in the ontology. Unlike codebooks, the ontology is not a list of definitions, descriptions, and examples, but a hierarchical representation of the relationship between the different terms.

In the seminar, several small groups worked on separate chapters regarding a specific topic that deals with the public discourse of AI. Since there was a common theme, i.e. AI in public discourse, around which all the chapters revolved, it is of importance that the course participants use a common ontology for their research. This ontology serves as a shared understanding of the field and its complexity and provides unique names for certain entities. Using the agreed names of the ontology makes it possible to compare different chapters and relate them to each other. For the analyses, the ontology serves as a tagset with which texts, or parts of texts, could be labeled. These tags help to examine a specific research question through, for example, analyzing tag combinations and frequencies. Overall, the shared ontology makes it possible to collaboratively work on a project as a whole course.

3 How did we build the ontology parts? - Description of the collaborative construction process

As part of an active, collaborative process involving the lecturer and the course participants, the ontology was elaborated. During the weekly course meetings, the lecturer specified thematic aspects of the ontology on which the participants had to work on together. These mentioned aspects will be outlined in the next section where the ontology will be thematised more precisely.

The general procedure for creating the ontology was as follows: During the seminar meetings, the participants were divided into smaller groups to discuss how the ontology should be structured with regard to the thematic aspect suggested by the lecturer. It was the group's task to discuss which points should be included in the ontology and how these points relate to each other. Most of the time, the groups designed mind maps to illustrate possible relations. Subsequently, the groups presented their ideas which were jointly discussed and elaborated by all course participants.

In the following sessions respectively, the lecturer summarized the findings of the various groups which were then again discussed in the course. Importantly, things could always be added and modified throughout the weeks such that the ontology was regarded as a dynamic course project. The drafting of the ontology occurred online and the course participants could comment on it and modify it to every time.

To create the ontology the lecturer proposed several preliminary fixings: First, the ontology should include hierarchies in the form of an "is-a" relationship. This requirement was implemented by illustrating the ontology as several mind maps. Furthermore, everything should be stated in natural language and the ontology should be independent from other already existing ontologies.

The ontology was developed before as well as during the students worked on their actual chapters. Thereby, in the beginning the process to obtain the ontology could be regarded as a top-down approach because the students used their already existing knowledge about AI and the discussion of AI in the public. After the students started to work on the chapters a revision of the ontology was still possible. This was a bottom-up approach because the new gained insights were used to modify the ontology. The usage of both approaches provided a broad perspective of the field which enhanced the formation of a versatile ontology.

4 What are the results?

To start off the process of building an ontology, two broad content points were proposed by the lecturer: “Agents and Places”. For Agents, the course decided to differentiate between “Agent Affiliation” and “Agent Experience”.

“Agent Affiliation” refers to the author or the agent communicating on Artificial Intelligence viz. a journalist/media, a business person/group, a governance body, etc. A hierarchy tree to represent Agent Affiliation was developed during the seminar and can be visualised in Figure1 to gain an insight on the discussed fragments. The hierarchy tree could be further elaborated for each of the nodes, although to keep it comprehensible and precise, only a few nodes have been branched to exemplify.

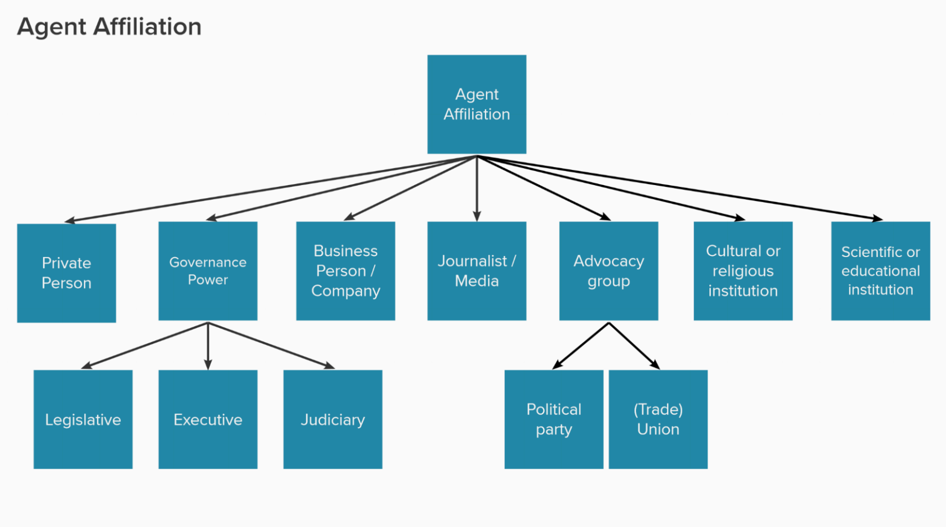


Figure 1: Agent Affiliation

“Agent Experience” distinguishes between the agents on the basis of their expertise in the field of AI. For example: If someone is a trained professional or simply an AI user, etc.

“Places” refers to the place of publication or the kind of media which is the source of the text being analysed. It is broadly differentiated between fictional and non-fictional texts. Fictional could be text from a movie script or a literature piece whereas non fictional sources could be the news, a blog post, a public speech or debate. Figure 2 illustrates a condensed hierarchy tree for the segments of publication being analysed.

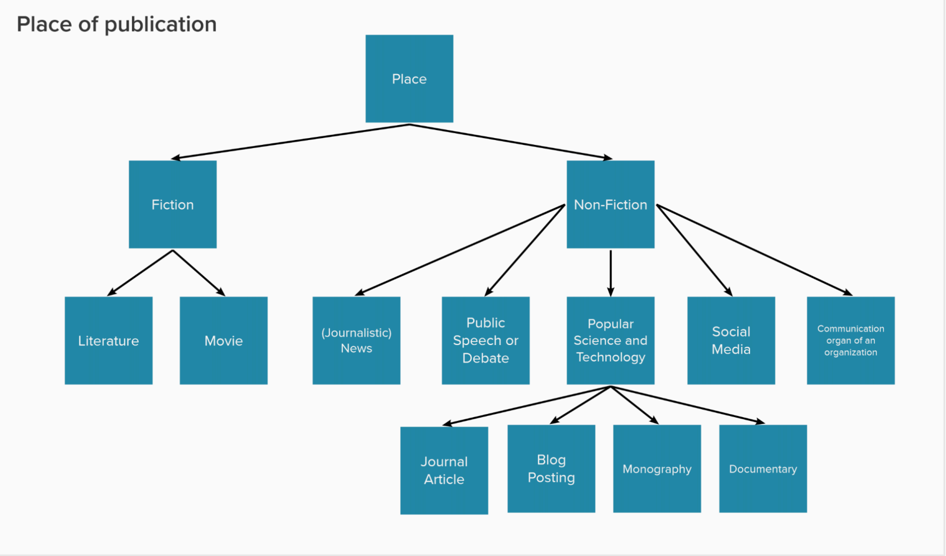


Figure 2: Place of Publication

While discussing Agent and Places, two important distinguishable entities to be considered emerged: “Date/ Time” and the “Type of the statement”.

“Date/Time” can refer to either “The date this statement was issued” viz. 2015-2021, 2000-2005, etc. or a “Time of Reference”, for instance: past, present time, near future (<5 years), medium-term future (5-20 years), etc.

“Type of statement” explores the intention of the author. For example it could be just factual reporting, an opinion or a text that intends to influence opinions.

The second content point proposed by the lecturer revolved around topics that were assumed to arise during the analysis of the public discourse on AI. The collected topics were divided into “Definitions of AI”, “Methods”, “Field of Application” and “Properties of AI systems”.

The section “Definitions of AI” focused on how AI is defined/perceived in the public. It was distinguished between “too broad definitions of AI” on one hand and the definitions that “reduced AI to a single aspect” on the other hand. For instance, a too broad definition of AI could be the idea of equating AI and digitalisation. An example of reducing AI to a single aspect could be defining AI as machine learning.

In the section “Methods”, AI was split into its sub-parts to distinguish which method of AI is referred to or represented in a text. Methods for instance are Machine Learning and Rule-Based systems, etc.

“Field(s) of Application” represents different sectors in which AI is applied. To exemplify some sectors here: Natural Language Processing, Health & Medicine, Mobility, Education, etc.

The section “Properties of AI” revolves around the characteristics of AI that are usually addressed in the public discourse viz. AI makes less mistakes than humans, has a high availability, etc.

“Attitudes, Emotions” cover several emotions that can possibly arise and consequently be expressed in texts in regards to AI such as hope, fear, doubt,etc.

The section “Anthropomorphisation” comprises topics that compares AI to human- beings or human-like qualities. To illustrate a few examples: “Machine Learning is similar to human learning” and sometimes even humanoid forms of AI, which are often gendered.

The section “Ethical Questions/Threats” comprises several ethical questions that could arise in the public discourse of AI, like accountability of decisions taken by AI algorithms, the transparency/explainability of the algorithms which act as a black box, bias in decision-making, etc.

The “Strengtgs, Weaknesses, Opportiunities and Threats” (SWOT) section of the ontology ensures to clearly define all factors influencing the analysis. Strengths and weaknesses can be seen as inherent properties of AI. Strengths describes everything that can show AI as an advantage whereas Weaknesses points in the opposite direction i.e. the areas of improvement. Opportunities and threats can be envisioned as the consequences and side effects, that result from the inherent properties, which can have either a positive or negative effect on society.

Lastly, the section “Proposal of Actions or Demand of Actions” includes demands on how to handle AI in order to cope with the negative consequences or ethical concerns that arise. Some examples of the proposed actions are “Legal AI regulations” ,fostering public discussion on AI, etc.

5 Full ontology / hierarchical list of terms and arguments

- Agent affiliation
    - Private Person
    - Governance Power (Legislative, Executive, Judiciary)
    - Business Person / Company

  - Journalist / Media
  - Advocay group (political party, trade union)
  - Cultural or religious institution
  - Scientific or educational institution
- Agent AI experience
  - Professional training / expertise (technological/computer science, ethical/philosophy, society/politics/law)
  - AI user (professional, academic, private)
  - AI developer
- Place of publication
  - Fiction (Literature, Movie)
  - Non-Fiction
    - (Journalistic) News
    - Public Speech or Debate
    - Popular Science and Technology (Journal Article, Blog Posting, Monography, Documentary)
    - Social Media
    - Communication organ of an organization
- Type of statement / intention
  - Factual report
  - Opinion
  - Intention to influence opinios
- Time
  - Date of statement (< 2000, 2000–2014, 2015–2021)
  - Time of reference (past, present time, near future (< 5 years), medium-term future (5–20 years), far future (> 20 years))
- AI Definitions
  - AI = The output of the scientific AI community in the last 65 years
  - Only string AI could be called real AI
  - AI = Software that tries to mimic human intelligence
  - AI = Going beyond what has been considered possible for computers so far
- Reduction to single aspects
  - AI = Machine Learning (AI=Neural Networks, AI Deep Learning)
  - AI = Robots (AI = Humamoid Robots)
  - AI is just advances statistics

- Too broad
  - AI = Digitalization
  - Any form of algorithmic decision is AI
  - Any dialogue-like user interface (chatbot) is AI
  - AI = Big Data
- AI methods
  - Machine Learning
  - Artificial Neural Networks
  - Deep Learning
  - Deep reinforcement learning
  - Knowledge representation
  - Rule-based systems
  - Automated reasoning
  - Planning
  - "Good old-fashioned AI" (GOFAI)
  - Virtual Reality
- Fields of Application
  - Mobility (Self-Driving Cars)
  - Security (AI in military, surveillance, predictive policing)
  - Health / Medicine (diagnosis, surgery, mental health support, service robots in health care)
  - Commerce (recommender systems, advertisment, credit scoring, stock/finance optimization)
  - Natural Language Processing (translation, conversational agents, text generation)
  - Household / service (lawnmower bots, cleaning bots)
  - Human Resources Departmens (job application decisions)
  - AI in science
  - Production
  - Education
  - Entertainment / Arts
- Properties of an AI system
  - might have emotions / can not have emotions
  - does not make mistakes / make less mistakes than humans
  - is inexplainable

    - weak vs. strong AI / is not “really” intelligent
    - cannot deal with unexpected situations
    - far faster than humans
    - high availability
- Attitudes, Emotions (hope, excitement, doubt, fear, disbelieve, euphoria, uncertainty, surprise, disappointment, boredom, anger, overwhelmed)
- Anthropomorphisation
    - AI is similar to human intelligence
    - Machine Learning is similar to human learning
    - “Artificial Neural Networks” are similar to the human brain
    - taking humans as a model for AI (physically and behaviorally)
    - talk about as an individual “We developed an AI”
    - humanoid forms of AI, often gendered
    - naming chatbots, giving them an identity and gender
    - transforming a human into an AI (fictional)
    - having emotions towards AI systems / robots
- Ethical questions / threats
    - Transparancy / explainbility / black box algorithms
    - accountability of decisions
    - problem of bias
    - privacy
    - reduction of human autonomy
- SWOT Analysis
    - Strengths
        - Scalability (always and everywhere, no breaks, can be copied and multiplied with no/little costs)
        - Complexity (can handle huge amounts of data, can find patterns that are too complex for humans)
        - Ability (can automate human intellectual tasks, con perform automated tasks very fast, can perform boring or repetive tasks)
        - Consistency (avoids typical human mistakes, reproducible results)
    - Weaknesses
        - Explainability (black box systems, very complex systems with unpredictable side effects, lack of understanding of AI in society)

    - Computationality (misses general understandind and context awareness, black box models do not fail gracfully, lack of emotions, vulnerable to biases)
    - Limitations (a lot of training data is needed, prone to errors because of noise and misbalanced data, limitations of capacity are unknown, some fundamental properties are not well understood, models are always limited)
    - High costs (high energy consumption, high expert knowledge needed, high computational power needed, access to extensive data sourvces needed)
  - Opportunities
    - Production increase (reduce cost and time, make tasks possible that would be too expensive or slow if done by humans, freeing time for people to do more creative jobs)
    - New kinds of problem solving (suitable way to solve unsolved problems, optimization of complex systems, assist humans in utilizing large amounts of data)
    - Higher quality (increases fairness by being objective, increases safety by making less mistakes, superhuman level of AI can be support for human flourishing, less work to do for humans)
    - Scientific advancement (increase understanding of human cognition by testing hypotheses, foster discussion about "new ethics")
  - Threats
    - Loss of control (deliverance to untrustworth non-understood systems, depedence on technology, application gains monetary value faster than understanding of it grows)
    - Amplification of negative tendencies (surveillance and privacy, can reproduce/intensify societal biases and prejudices, power concentration, could be used as a means for ever increasing growth amplyfying problems like the climate crisis)
    - Destruction (AI kills human jobs, too high consumption of resources, anti-democratic (taking huamsn out of the loop), new vulnerabilities to be exploited by hackers and malware)
    - Lack of Acceptance (fear of humans because of not knowing enough about AI, loff of trust in economic and political system)
- Proposal of actions of demand of actions
  - no actions necessary
  - Legal AI regulations
  - Ethical guidelines

- start/foster public discussion
- counteract power concentration (e.g. by anti-monoploy measure, or by funding European projects)
- moratorium
- social and economical restructuring and reformation
- Information / education about AI
- increase investment / research in AI

References

Gibbs, Graham R. 2018. *Analyzing qualitative data*. Vol. 6. Sage. Chap. 4.

Lapan, Stephen D, Marylynn T Quartaroli & Frances J Riemer. 2012. *Qualitative research: An introduction to methods and designs*. Jossey-Bass. 42, 43.

Leavy, Patricia. 2014. *The Oxford handbook of qualitative research*. Oxford University Press, USA. Chap. 1.

Vaismoradi, Mojtaba & Sherrill Snelgrove. 2019. Theme in qualitative content analysis and thematic analysis. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, vol. 20.

Part I

The general public

Chapter 3

How do non-experts discuss AI?

Michael Alexandrovsky, Mekselina Doganc, Virginia Jagusch, Julia Stormborn & Daniel Pietschke

This chapter addresses the question of how private persons with no extensive knowledge about Artificial Intelligence (AI), who in sum make up the general public, talk about AI. In order to adequately maneuver through the social challenges related to AI, it is of major interest for policy and decision makers to understand how these non-experts perceive, discuss and reason about AI. So far, no extensive research has been conducted to capture the perception of non-experts on AI directly. Previous research tried to capture the public discourse primarily from a media or expert-driven perspective. This paper aims to close this gap and present some qualitative, albeit not representative, findings from four non-expert discussions about AI. The discussions were undertaken in groups of 3-4 German-speaking young adults. The subsequent analysis examines which topics the participants discussed, which emotions they expressed, and what action potential (i.e. actions that can be taken to contribute meaningful change to society) they perceived concerning AI. According to our qualitative analysis, the participating non-experts speak about the following topics when discussing AI: Self-driving cars, automation, privacy, social media and big data, and lack of education and information. Expressed emotions were mostly fear and worry-related. The participants perceived only a very limited or no personal action potential, however, they raised many ideas for taking action on a societal level.

Keywords: Artificial Intelligence | Public Discourse | Discussion Groups | Public Opinion | Autonomous Vehicles | Automation

1 Introduction

As research and development of Artificial Intelligence (AI) progress, so does its march from academia into the real world and our everyday lives. People are con-

All authors contributed equally. The authors are alphabetically ordered.

stantly confronted with new technologies that use some sort of AI, be it in their smartphones, photo editing software, or washing machines. This points to a future in which AI solutions and applications will become even more important than today and will affect everybody. If this is true and AI will have such a huge impact on humanity, it is essential to open the discourse for non-experts and find out find out their opinion and their reasons for it. Today's society already has to face many ethical issues related to AI and its applications, such as face-recognition of people of color (Buolamwini & Gebru 2018), decisions related to imprisonment(McKay 2020), discriminating application procedures (Dastin 2018) and more. Addressing these issues should be included in a democratic decision-making process. A positive example can be found in a survey conducted by the European Union to consider the public opinion on AI (EU Commission and others 2020).

In this paper, we define 'Non-experts' as persons who do not work with AI in an academic or professional context. Finding out what non-experts think about AI and understanding their reasoning can be considered a first step towards guaranteeing public participation in shaping the future of AI decision and policy-making.

This paper approaches how non-experts perceive AI by addressing the following questions:

- Topics: What were the discussed topics related to AI, what standpoints are held and for which reasons?
- Emotions: Which emotions did people express, when talking about AI?
- Action potential: In which AI-related areas do people see the need to take action?

To our knowledge, only little research on this has been done in the past. Fast and Horvitz performed an analysis of thirty years of New York Times articles to find out how the public perception, attitude and topics about AI have changed over time (Fast & Horvitz 2017). They discovered that since 2009 (following the popularisation of deep neural networks) the subject of AI has become more and more prevalent. At the time of the study (2017), the most common topic was autonomous vehicles. The absolute amount of optimistic as well as pessimistic articles has gone up. Usual concerns are about the loss of control over an AI, how to deal with ethical problems and how AI will impact the job market. Hopes are expressed particularly towards applications of AI in healthcare. In 2020, the European Union issued a public consultation on their white paper on the European approach for developing AI (EU Commission and others 2020). They received over a thousand responses from different agents in society, such as citizens, businesses, industry and academia. In their evaluation, they found that particularly civil society sees great importance in 'democratic oversight' and is 'sceptical of self-regulation' of that matter.

In this chapter, we investigate four discussions regarding AI held by unprepared non-experts. We assume our investigation will provide valuable insights into which positions are commonly held by non-experts, which assumptions these are based on, and which emotions are most prominently associated with AI. We hope that these insights will help to facilitate communication between the developers of AI, policy makers and the end-user in order to make the topic more accessible to the wider public. The discussion approach was chosen since discussions in peer groups are important social means for exchanging information and forming, reinforcing, or changing the group members' opinions.

We begin by describing our research methods which consist of organizing and carrying out discussions of non-experts as well as the analysis of the discussions. We proceed with the description of our findings which we will divide into the three parts *topics, emotions and action potential* as described above.

We will conclude with a discussion on our procedure, our results and suggestions for further research.

2 Methods

In this section, we first familiarize you with the data we have collected and used. Secondly, we will present the discussion procedure and finally explain how we analyzed the collected data.

2.1 Study Participants

We organized four group discussions between non-experts in the field of AI. We defined 'non-experts' as persons that do not work with AI in an academic or professional context. The total number of people who participated in the main discussions was 15 (8 female / 7 male, aged between 21 and 25). The participants were selected manually by us to be confident speakers in German and to be likely to engage in discussions with strangers. We split the participants into four groups of 3 to 4 people. With the consent of all of the participants the discussions were recorded via Open Broadcaster Software for further analysis. The participants were informed that the recordings will be deleted after the final submission of this chapter.

2.2 Discussion Procedure

The discussions were held via the online conference platform Big Blue Button[1] which could be accessed in an online browser via an invitation link. All participants were asked to ensure a stable internet connection, to be in a non-distracting environment for the time of the discussion, and to enable their front cameras. We expected the discussions to be about one hour long. The participants came unprepared to the discussions and were informed about the topic only at the beginning of each discussion.

Additionally to the four main discussions, we also organized one preceding mock discussion with four additional participants. It was crucial to test the setup and it allowed us to refine the structure and reformulate some of the questions, e.g. to not bias the participants. The mock discussion was not included in the final evaluation.

The discussions were divided into three parts. The first part consisted of a brief introduction with the announcement of the discussion topic, the participants stating their names, age and occupation, as well as giving consent to be video and audio recorded. The participants also shared their understanding of what Artificial Intelligence is.

The main part was the discussion itself. We provided three sets of questions which are listed below:

Block I

- Where do you encounter Artificial Intelligence in everyday life?
- Where do you see the benefits of Artificial Intelligence?
- Where do you see the downsides of Artificial Intelligence?
- What kind of influence does Artificial Intelligence have on society?
- Who is affected by Artificial Intelligence?

Block II

- Do you think Artificial Intelligence is ‘good’ or ‘evil’? (And why?)
- Where do you see action potential?
- Where do you see personal action potential?

[1] Hosted on the servers of the Osnabrück University.

Block III

- Can you agree upon a definition of Artificial Intelligence?

Each of these question blocks was revealed and handled separately before moving on to the next block. The questions within a block were presented all at once on a single slide.

To ensure a natural conversation and to remove the fear of providing wrong answers (which would have been detrimental for our analysis), we told the participants that there were no wrong answers to these questions. The questions were presented to start a conversation and the participants were allowed to digress and let the discussion flow freely.

During the discussion, there was almost no moderation from our side to avoid biasing the participants. We only got involved when the participants told us that they are ready for the next question block, when we had the impression that we had to move on due to time constraints, or the participants strongly deviated from the topic of Artificial Intelligence.

The third part consisted of debriefing the participants about the purpose of the study and a voluntary after-discussion with the moderators, where we answered questions that came up during the main discussion.

2.3 Analysis

The analysis consisted of two parts: A quantitative and a qualitative approach. For the quantitative approach, we identified topics and categories using the established ontology. The identified categories were *Definitions, Emotions, Ethical Questions, Field of Application, Opportunities, Threats* and *Properties*. We looked at how frequently elements from those categories were explicitly mentioned by the participants, noting any significant majorities or tendencies towards a certain category. The results of the quantitative analysis regarding the categories *Emotions, Ethical Questions, Field of Application, Opportunities* and *Threats* were used to support the qualitative analysis as well. For the qualitative part of the analysis, we identified central topics the participants discussed in-depth, alongside the detailed arguments made by participants to discuss these topics.

3 Results

3.1 Topics

As the participants were asked about what they believe the advantages and disadvantages of AI are, several themes and arguments could be identified.

3.1.1 Self Driving Cars

One major topic was the discussion of self-driving or autonomous cars. This topic was identified as recent and relevant to the current state of AI and led to an in-depth discussion, with several arguments re-occurring over and over during each discussion. As an advantage of self-driving cars, participants noted an increase in comfort in daily life, as well as the reduction of accidents involving self-driving cars. Participants argued self-driving cars will likely outperform human drivers. Some participants noted they would gladly use self-driving cars and seemed excited about the idea, saying they would believe logic and data leading to those cars to be safe. Other participants voiced discomfort with autonomous cars and argued that they would still feel unsafe and insecure, regardless of any data proving autonomous cars to be safe. The reasons for these hesitations varied. Some participants noted they would not like to give up their autonomy on the road, and would especially like to retain control of the car in emergencies. A fear of external parties 'hacking' the car and taking control of it was noted as well. Participants questioned whether the AI of an autonomous car would be able to make split-second ethical decisions when faced with a dilemma. One participant in particular argued such ethical dilemmas would not even occur, as „high-tech cameras all around the car would foresee such circumstances" and be good at preventing all accidents.

3.1.2 Automation

Another major topic was automation, or the use of robots to replace human workers in certain fields. Participants noted that automation would be cheaper and more efficient, and might help eliminate repetitive or inhumane work; as well as to fill out 'unpopular jobs', such as the understaffed medical field. Upon this, the question was raised what would happen to the people who lost their work in the process of automation. The suggested solutions ranged from temporary financial compensation of laid-off workers to universal and unconditional basic income for all humans. Further questions were raised by participants asking when there may be an explicit need for a field to become more modern. The cases of workers unwilling to lay down their jobs because they found fulfilment in them were mentioned; or fields of work that would not necessarily need AI, but where it would be included because of the hype surrounding it. Another related issue was raised by the participants as they noted that the shift in the workforce would benefit highly specialized, academic workers more than 'more practically oriented' people, who would have trouble keeping up with the development and may be outperformed by machines. In general the question of ’getting left behind’ was raised. Participants noted on several occasions that they themselves, and many of their peers, lacked the technical proficiency necessary to understand, or even want to understand, the algorithms behind artificial intelligence, and expressed worry over this lack of knowledge.

3.1.3 Privacy, Social Media and Big Data

When talking about ethical issues surrounding privacy and social media, the discussion veered off faster into different topics compared to the previous topics mentioned. Participants differentiated little between issues of digitalization in general and issues directly related to AI applications. In this context, they wondered when an algorithm could be considered intelligent as opposed to a regular algorithm. Nevertheless, privacy was an issue raised in all of the discussion groups. They noted the lack of transparency of both the algorithms and the exact usage of their personal data, referencing social media advertising and suggestion algorithms as well as smart-home applications in particular. Some participants expressed the worry that data collecting AIs are so deeply nested into society, that it would be impossible to refuse using them and to give them private information without isolating oneself from society. It was also noted that this data gathering would be already unavoidable, with people’s information getting picked up anyway even if they refused to use social media, by virtue of being near humans who kept using it and getting recorded by their devices instead. Participants also mentioned the role social media in particular plays with (political) radicalisation, claiming that the ’suggestion algorithms’ driving most social media platforms trapped people in a sort of „bubble“ or „echo chamber“, where they would only be exposed to one certain opinion, isolating them from a variety of worldviews. Several participants also noted that, while annoyed with the misuse of their data, they would not like to give up the usage of these AI applications, and would just accept the non-transparent data collection.

3.1.4 Lack of Education and Information

Another disadvantage noted was the lack of widespread formal education on AI and its applications. The Participants believe the public mistrusts AI on principle, which they explain by the little general understanding of modern technologies. The lack of formal education in schools was pointed out and participants expressed the wish for students to be taught more extensively about AI and of a higher standard. Several participants also claimed to be glad that students these days already possessed more possibilities to learn about these things than they did as students, for example by having more detailed computer sciences classes in school. Beyond school education, the elderly should not be excluded from such educational opportunities as well. The lack of official and formal sources to reliably inform oneself about AI was also an issue raised by the participants, as they mentioned that widespread misinformation about the topic hindered them in educating themselves properly. Another topic raised was the belief that politicians may be just as uninformed as the general public on the issues of AI, and should correspond more with experts in the field. One participant in particular noted that they don't expect politicians to be able to operate a Facebook account, much less understand complex AI algorithms.

3.2 Emotions

When speaking about how they feel about AI, participants predominantly explicitly expressed the emotions 'fear' and 'worry'. 'Explicitly expressed' in this case means participants mentioning their feelings directly as „I feel X“ or „this makes me feel X“ or similar phrasings where the emotion expressed is abundantly clear and not dependent on any interpretations of, for example, body language. 46,6% percent of the total mentions were expressions of fear, the predominant word used being „scary“. The contexts or aspect of AI which participants mentioned being afraid of varied, and could be found in any of the topics discussed above. Furthermore, participants expressed the emotions of 'worry' and 'doubt' a lot. Positive emotions such as curiosity or excitement towards AI and its future only made up 10% of the explicit mentions of emotions in the discussions. [Participants heavily focused on disadvantages and controversies during their discussions and positive aspects of AI tended to be listed briefly or met with negative rebuttals.]

3.3 Action Potential

3.3.1 Personal Action Potential

When answering the question about personal action potential, participants expressed a lot of uncertainty. Several participants either didn't see any personal action potential necessary or believed they, as singular people, would not be able to have any kind of impact anyway. When discussing action potential in terms of privacy issues, participants discussed whether self-exclusion from social media to avoid the invasions of their privacy would be a benefactory action. They found they would not be able to escape the selling of their data in the long run or would be unwilling to give up the functionalities of social media, tolerating the perceived invasion of their privacy. The actions which participants proposed most often were being aware and informed about AI and its issues. They would like to use AI applications more mindfully, to be aware of their advantages and drawbacks, while also saying to inform themselves more on these issues and the exact, more detailed workings of AI and its' applications. Other participants, nevertheless, questioned whether this would lead to them trusting AI applications more, or whether they would be able to properly educate themselves on these topics in the first place.

3.3.2 Societal Action Potentials

When discussing societal action potentials, participants' answers varied a lot more. In general, they wished for more discussion and dialogue around AI in public, especially from politicians, as well as the provision of information on AI in public schools. One suggestion was that politicians should potentially already start considering laws and rights for what would we consider strong AI, and especially further privacy laws to prevent companies from over-collecting and misusing data. In terms of particularly concrete solutions, one participant noted the lack of consensus on where AI can and should be used, and in which fields it could not be applied without serious societal drawbacks, and wished further societal dialogue on these topics. Another participant suggested the implementation of an „internet-tax", where users would pay a small fee per google-search, to raise mindfulness for the use of smart technologies. A third participant proposed the implementation of the unconditional basic income for humans to ensure the quality of life during the shift to full automation, as well as severance packages for people affected by the abolishment of their work fields.

4 Reflection and Limitations

When designing the questions we tried to phrase them in a way which does not bias the discussion members to focus on certain aspects of AI. However, our analysis shows that some topics were discussed extensively (self-driving cars, automation etc.). Other topics, which we expected to fit into the common understanding of AI (particularly form a *science-fiction* perspective known from films and books, such as superintelligence, machine consciousness, machine rights), were either mentioned only briefly or not discussed at all. A possible explanation might be that our questions did bias the participants. Perhaps the question *Where do you encounter AI in everday life?* implied that we expected the notion of AI to be based on the current and 'near future' applications of AI. It is also possible that this assumption may be drawn from the scientific context of the discussions and the participants thought that we would not be interested in a more ’futuristic’ notion of AI. Other factors, such as time restriction, peer pressure, homogeneous groups or how confident people felt talking about a topic might also account for the similarity of the discussed topics across the discussion. Perhaps these are simply the topics that *are* usually discussed by non-experts. An answer to this question might be found in a following rigorous investigation.

We didn’t pay attention to the frequency a topic was mentioned, and did not consider how many participants exactly agreed with which sentiment. The first one was neglected as we wanted to focus on the in-depth arguments and opinions of the participants, and not whether they tangentially remarked on a topic without discussing it. For similar reasons, it wasn’t important that all participants agreed with each other, as our number of participants is not representative anyway, we focused on the dialogue. In many cases, participants expressed uncertainty about their knowledge and opinions or made vague statements, which were hard to consider for quantitative analysis, which was another reason for neglecting this approach in the results.

Our quantitative approach to the data using the categories of the ontology turned out to be inappropriate for the data in several cases. The prompts given by the categories *Definitions* and *Properties* of the ontology either were not mentioned by participants at all, or their mention didn’t give any insight to the participants opinions. As our results of the quantitative analysis are not representative due to the sample size of participants, we elected to not include them in the results sections of this paper. The data offered no significant our outstanding insights on these accounts and remained inconclusive. The only exception is the *Emotions* category, which did provide significant results which were listed in the previous section. Although the prevalence of emotions related to fear and doubt stood out, no clear or direct reason for the prevalence of these emotions could be identified.

5 Conclusion

The main finding of this paper is that analysing non-experts discussions on AI is an informative and feasible method to address the question of how non-experts talk about AI.

When discussing AI-related topics, participants across different groups tend to discuss similar topics. Those topics are not necessarily solely AI-related and more commonly they are in the bigger scope of digitalization. Especially, when discussing topics related to privacy, social media, and big data the discussion often wandered from AI into other fields. This suggests that non-experts do not necessarily separate topics such as AI and digitalization clearly from each other and interlace different tech-related fields when discussing remotely AI-related topics. Self-driving cars and automation were discussed intensively. The different arguments concerning self-driving cars lead some to an affirmative position and others to a negative position towards self-driving cars. In contrast, automation was discussed more consistently from a negative point of view. The participants often expressed their perceived personal lack of knowledge about AI. Those expressions are consistent with the wishes expressed while discussing the lack of education and information on AI for the general public. The participants wished for better general education on AI in schools and official, publicly available sources providing information on AI.

The emotions explicitly expressed during the four discussions were mostly fear, worry, and doubt, with little mentions of positive emotions such as curiosity. Those findings do not completely mirror the reasoning observed in the discussions which also expressed more positive positions towards specific AI applications. However, the positive emotions related to these positions were expressed more implicitly.

Participants assessed the action potential regarding AI very differently, depending on whether they had been asked for their personal action potential or the societal action potential. In general, they perceived little or no possibility for personal action potential because their actions would have no impact on society. However, societal action potential was viewed very differently: Many different actions were suggested, such as initiating more public discourse, starting on lawmaking for AI, or introducing an unconditional basic income to ensure human-well being in the face of automation.

Future research could analyse if those topics and perceptions reoccur when observing more than four discussions with a wider demographic scope. The discussions could be repeated with more diverse participants regarding age, occupation, social situation, and cultural background. It could also be of major interest to find ways to quantify research on public AI perception in order to complement the qualitative findings presented in this paper with figures.

All those findings could be essential for decision and policymakers since laws and policies on AI find only support from the general public if those topics are addressed. Policy and decision-makers also need to be aware of the emotions that non-experts associate with AI in order to address their fears and worries. If the impact of AI on our society should take place in a democratic process, the ideas of non-experts should be heard to mitigate negative social impacts from AI applications.

References

Buolamwini, Joy & Timnit Gebru. 2018. Gender shades: intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler & Christo Wilson (eds.), *Proceedings of the 1st conference on fairness, accountability and transparency*, vol. 81 (Proceedings of Machine Learning Research), 77–91. New York, NY, USA: PMLR. http://proceedings.mlr.press/v81/buolamwini18a.html.

Dastin, Jeffrey. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (31 March, 2021).

EU Commission and others. 2020. White paper on artificial intelligence — A European approach to excellence and trust. *COM (2020)* 65.

Fast, Ethan & Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence* 31(1). https://ojs.aaai.org/index.php/AAAI/article/view/10635.

McKay, Carolyn. 2020. Predicting risk in criminal procedure: actuarial tools, algorithms, AI and judicial decision-making. *Current Issues in Criminal Justice* 32(1). 22–39. DOI: 10.1080/10345329.2019.1658694.

Chapter 4

What impact have experts on the public discourse about AI? - An analysis of reactions to newspaper articles on “ZEIT-Online”

Sascha L. Mühlinghaus

This work analyzes reactions to public media contributions from experts for artificial intelligence. The public discourse about artificial Intelligence is contradictory due to many different people who contribute to it. Therefore it is required that experts give an overview and context to certain positions. In the following, the comments of two expert contributions were analyzed and categorized systematically based on criteria that evaluate content-related and formal aspects. The dataset consisting of comments was generated from articles on the German newspaper platform ”Zeit-Online”. This paper expands on the perspective of resonance to expert contributions by the public about attitudes and evaluation of AI systems. Many commentators had a rather critical opinion on the author’s statements and criticized single aspects or argued towards undermining the credibility of the expert. Further, the abilities of artificially learning systems were often assessed as not capable, about half of the comments noticed a certain suspicion regarding AI.

Keywords: AI | AI-expert | Zeit-Online | commentator | reaction | public discourse

1 Introduction

Experts from the field of AI (artificial intelligence) who state their view on AI in mainstream media seemed to be rare during this research. Although the interest in AI on public media is increasingly immense (Alsheibani et al. 2018).

The field of applications for AI systems is very broad and complex which requires expertise to understand. The high anticipation of the matter leads to a wide range of contributions from many different platforms with contradicting views, further the expertise varies strongly among these contributions. In such a situation it is important that experts raise their voice as an authority and substantiate the discourse with specific knowledge. In this paper, it is investigated how contributions from people who are involved in AI professionally are discussed and perceived by the public. Hypothesizing that experts from varying disciplines are perceived and evaluated differently by the public. The analysis focuses predominantly on the respective reactions in the form of comments and partly also on the expert-contribution itself.

The two extracted contributions are the essay ”Roboter können keine Moral” (Robots have no morals) by Richard David Precht (Precht 2020) ”Lassen wir die Finger davon” (Let´s keep our fingers off) where the guest Oliver Bendel is interviewed by Georg Etscheit (Bendel & Etscheit 2017).

These articles were selected from the German newspaper platform ”Zeit-Online”, the platform also offers the opportunity to comment and to participate in the discourse about published articles but only as a registered member(free of charge). By analyzing the immediate reaction of the readers, the setting allows to look at possible impacts the authors had on the audience. Concretely, what the main focus of this analysis is, are differences in the evaluation of AI depending on the representation of the author. Further, this approach might be an opportunity to assess the importance of experts taking part in a public discourse to increase the public acceptance of AI technologies.

2 Methodology

The qualitative analysis performed was supposed to highlight and categorize differences and similarities among the comments. This was done in order to give an overview of the spectrum of reactions to expert contributions on the discourse about AI. A similar analysis is applied to the articles extracted but with the intention of giving a background and possible explanation for the reaction of the reader.

2.1 Data Selection

The two sources that were extracted meet the requirements of presenting the unmediated opinion of an expert of AI. In this paper, an expert of AI is liberally defined as a person that labours in academia and conducts work in the field of AI. Principally, this approach not only defines computer scientists or AI programmers as experts but also people who are involved with other disciplines that focus on ethical or societal characteristics that matter to AI. Further, the contributions should address a broad, not necessarily academic public. Besides, as a reader, the possibility to publish comments was crucial for this research. Moreover, the sources were chosen because the respective experts did not only describe the current state of the art in the field of AI but also provided a perspective on possible developments. The personal assumption made based on the research about this aspect is that such contributions would evoke more comments on the articles.

2.2 Method Application for the Comments

The following are the areas of interest for the comments that were systematically examined.

- What point of view or prior bias towards AI and the contribution does the comment reveal?
- Were technical terms used and if so, were they used correctly? This aspect was evaluated considering the ontology created in the seminar course ”AI in public discourse” (cf. chapter 2). The intention is to have an idea about the technical knowledge of the reader.
- What purpose was the commentator pursuing?
- Is the comment rather technically or emotionally formulated and argumentative?
- Is the comment apparent to be optimistic or suspicious regarding the development of AI?

2.3 Method Application for the Articles

As the articles are not the main focus of this investigation, only a superficial analysis will be applied to the expert contribution. This is meant to give the comments an embedding and to possibly draw connecting conclusions from the features of the articles and the comments.

Therefore, a short and concise summary of the author’s view is given, furthermore the use of technical terms in the article will be considered, regarding the ontology of the course 2. The two authors are introduced with information about their specialization and education. At last, the articles are analyzed regarding their focus on future or present aspects.

3 Analysis

As the inspected data is collected from only few sources the investigation does not demand to be representative. However, in order to reveal possible biases prior to the analysis, information about the newspaper platform and the experts is provided.

3.1 Zeit-Online

Zeit-Online is a medium that is partly freely accessible, the article of Richard David Precht published in 2020 belongs to the "Zeitplus" section which requires a chargeable subscription. In contrast the article from Oliver Bendel published in 2017 was published on "Zeit-Campus" (URL: https://www.zeit.de/campus/index, which is a section that aims at young academic people and is freely accessible. Zeit-Online is a well known online portal in Germany, the editor-in-chief Giovanni di Lorenzo describes "Die Zeit" as a "left wing" liberal newspaper (di Lorenzo 2021). The readership of Zeit-Online is not representative for the broad German public, according to the readership data the education of Zeit-Online readers is on average higher than among all German citizens (Goedert n.d.). Nevertheless, Zeit-Online is a highly frequented media platform, despite the fact that the readership of Zeit-Online represents a rather scientifically interested audience. Additionally, due to the fact that most readers do not leave a comment, the picture that is drawn by this analysis represents a special niche of the readership.

3.2 Oliver Bendel

Oliver Bendel is a professor for business informatics at the University association Northwest Switzerland and is a scientist with a focus on machine ethics (*Der Wissenschaftler Prof. Dr. Oliver Bendel* n.d.).

Oliver Bendel is interviewed in the article "Lassen wir die Finger davon" (Let´s keep our fingers off) by Georg Etscheit published 2017. In the first part of the article, the questions address projects of Bendel about lying simulations on the internet. The questions address mainly ethical issues in the context of AI. In the following, Bendel is talking about upcoming issues with self driving cars (Bendel & Etscheit 2017). When looking at the reactions, the content of the article is examined in more detail, if necessary.

3.3 Richard David Precht

In Germany Richard David Precht is a well known Philosopher and writer, who published various books about philosophical as well as societal subjects. He also holds professorships for philosophy at the Leuphana University Lüneburg (Zühlsdorff n.d.) and at the University for Music Hans Eisler in Berlin (*Prof. Dr. Richard David Precht* n.d.).

The article ”Roboter können keine Moral” (Robots have no morals), as the title suggests, mainly speaks about AI and its incapability of acting ethically. The article is based on Precht’s 2020 book publication, ”Künstliche Intelligenz und der Sinn des Lebens” (Artificial Intelligence and the meaning of life) and indicates that a human differs in principle in its cognitive constitution from a computer. Consequently, according to Precht that makes it impossible for computers to simulate what is necessary to make ethical decisions (Precht 2020).

3.4 Comparison of the Articles

In both articles similar examples of AI appear but Precht deals with a general question about the capability of AI whereas Bendel is looking at current developments in the year 2017. The articles appeared with a temporal difference of about three years and represent positions that disagree in certain points. It offers the opportunity to compare reactions of the reader to aspects mentioned by both authors in the respective articles. Precht argues about the ability of AI to act ethically and make moral decisions (Precht 2020). This application of AI refers to a situation where AI is forced to act autonomously. Also this action and previous calculation on how to act is highly ethically relevant. In my opinion such an autonomous ethical decision turns this case into a possible future application of AI that is not yet established.

It seems reasonable to assume that thoughts about the future of AI are more theoretical and even speculative than speaking about AI systems that have already been tested and developed. In contrast, Bendel focuses on present and presumably near future aspects of AI, like self-driving cars and bots that interact with a virtual environment. Bendel himself was involved in an online application of a ”lying bot”, he says: ”Ich will zeigen, dass es ohne Weiteres möglich ist, moralische und unmoralische Maschinen zu konstruieren […]” (I want to show that it is easily possible to construct moral and immoral machines) (Bendel & Etscheit 2017). Precht engages in a more philosophical position about what makes humans different from machines. He argues that morality is always connected to feelings and values which can not be described precisely nor consistently due to the dynamic of human emotion. Therefore, these processes cannot be simulated by a computer (Precht 2020).

The interviewer Etscheit asks about current fields in which AI or semi-autonomous systems are being tested to which Bendel gives some context of ethical issues, e. g. the technical issue of hostile/civilian recognition in autonomous weapon systems. In the further course, Bendel is also asked questions about future technologies. He anticipates problems but is not definitive in the decision about whether a technology can be applied in a certain position or not(Bendel & Etscheit 2017).

3.5 Comment Evaluation

In this section, it will be investigated how the comments appear under the aspects named in 2.2. Further, it was intended to connect the observations made to content related features of the respective articles.

In total there were 53 comments on the article ”Roboter können keine Moral” and 97 comments on ”Lassen wir die Finger davon”, 25 and 21 comments referred directly to the respective article. The remaining ones were either reactions to other comments or not utilizable due to a lack of content and therefore excluded.

In general, among the comments to Precht‘s article, about half of those reactions criticized Precht‘s ability to evaluate the technical component of AI or spoke against certain aspects of his argumentation. Like the commentator ”karimragab” writes:

> ”Trotzdem geht seine Argumentation weit am Thema vorbei. Statt sich an Philosophen der Vergangenheit zu orientieren, hätte er besser die Wissenschaft der künstlichen Intelligenz ein wenig tiefer studieren sollen.” (Nevertheless, his argument goes far beyond the topic. Instead of looking at philosophers of the past, he should have studied the science of artificial intelligence a little deeper) (Precht 2020).

Another user ”CY007” states: ”Ich glaube Precht hat von IT und im speziellen KI keine Ahnung[...]” (I think Precht has no idea about IT and especially AI) (Precht 2020).

This criticism might be ascribed to the fact that Precht is involved in many public discussions. As a philosopher, he is not only specialized in the field of AI. On the other hand, another Zeit-Online article criticized AI programmers for having too little ethical knowledge, which a philosopher should have (Wolfangel 2021). Precht might be evaluating the matter from a different point of view.

Compared to Bendel, five comments addressed the article involving criticism, Precht faces eleven comments. This observation can probably only be explained by a multi causal relationship. Among these causes is the time difference of three years in which the (pseudo-)knowledge of the public about AI increased and consequently the confidence to react critically. Further, it could be due to more daring anticipations about the capabilities of AI by the author. Also the number of comments that had a rather suspicious perspective on AI is lower for the commentators of Precht‘s article. This observation probably corresponds to the time shift. At last of course his opinion can be less accepted publicly which could have caused the criticism.

Bendel is often criticized for his opinion to limit autonomously driving cars to highways and prohibit this feature in cities. He argues that the dilemma of a self driving car that can not avoid an accident and has to weigh up the harm of human lives is inconvenient. Many commentators state contradicting opinions on this standpoint e.g. by ”CEPTION”:

> Mir erscheint es da am moralisch sinnvollsten, die möglichen Personenschäden aller an der Situation beteiligten(sic.), im und außerhalb des Fahrzeugs, aufzuaddieren. (It seems to me that it makes the most moral sense to add up the possible personal injury to everyone involved in the situation, inside and outside the vehicle)(Bendel & Etscheit 2017)

Or ”AZ” (Bendel & Etscheit 2017): ”[...]dass Maschinen in Zukunft viel sicherer fahren werden als Menschen es tun und können; es wird sich so also viel menschliches Leid vermeiden lassen” (machines will drive much safer than humans can and do, so much human suffering can be avoided in this way). It appears that Bendel is provoking a concrete debate about specific aspects and is therefore criticized for certain positions but the audience is less likely to illicit Bendel as an expert.

Looking at how AI was described in the comments, I could observe that among the comments to Precht’s article more commentators had too broad or incorrect definitions of AI, this was evaluated regarding the definition given by the ontology of the course. The commentator ”JWGRU” (Precht 2020) brings up ”das kann auch ein Computer oder, wie das heute genannt wird, die künstliche Intelligenz” (a computer can do that too, or as it is called today, artificial intelligence). It appears that in this case, the distinction between AI and conventional computer programs is missing. This undifferentiated explanation occurs also in Precht’s (Precht 2020) argumentation ”Der Begriff ”künstliche Intelligenz” wirkt auf viele IT-Experten wie ein Marketingtrick” (The term ”artificial intelligence” appears like a marketing ploy to many IT experts). What he might want to say is that AI is sometimes overestimated beyond its capability to learn and to solve specific tasks. Precht does not make the differentiation clear between strong and weak AI. Highlighting this, the difference between conventional computer programs and AI systems is not clearly defined by Precht.

This may be reflected in more comments to the essay ”Roboter können keine Moral” (Robots have no morals) that exhibit incorrect use of technical terms (six comments) than to Bendel‘s article (four comments). Among the reactions where technical terms are used correctly, there are seven out of twenty-five correct for Precht and ten out of twenty-four correct for Bendel respectively, in other comments, no technical terms were used. Bendel specifies the difference between self learning systems and his own ”lying bot” which was not labeled as AI: “Der Microsoft-Bot war, im Gegensatz zu meinem Lügenbot, ein selbstlernendes System.” (In contrast to my lying bot, the Microsoft bot was a self-learning system) (Bendel & Etscheit 2017).

The categorization of whether the comments showed a suspicion or optimism regarding the development of AI, indicated less clear reactions to the essay by Precht than to the interview with Bendel. Fifteen comments to Precht did not reveal an exact standpoint. Seven expressed suspicion and two optimism. In 2017 in response to Bendel’s interview, eleven showed suspicion, four optimism and six no identifiable position. I am not sure whether this aspect can be placed as support for any of the above given hypothesis, since it is most likely an effect of the time that passed between both article appearances. It indicates that in 2017 the public was more pessimistic about how AI would influence our lives.

4 Results and Outlook

Two aspects are highlighted, the first one is that in 2017 Bendel addressed aspects of AI that are concrete and differentiated developments that affect our near future and are fixed to a specific application like autonomously driving cars. The interviewed Bendel comes from a subject that is closely related to computer science, he afterwards specialized in the field of machine ethics (*Der Wissenschaftler Prof. Dr. Oliver Bendel* n.d.).

In his essay from 2020 Precht focuses on a general perspective about the capability of AI to make ethical decisions. He is a philosopher and a person who is involved in many public discussions but nevertheless an expert in the field of philosophy and a driver of societal debates. Precht legitimates his presence in various discussions because he experiences himself as a connector of different disciplines and therefore looks at a more general spectrum of AI.

As a whole, we can see that the broad frame of Precht's approach led to a more conflicting discussion among his readers. Compared to a more specific approach of Bendel, that might have led his reader to more depth in certain aspects of the topic. We could see that the two different approaches chosen by the experts were portrayed in different reactions by the reader. Probably, both kinds of approaches are required in order to keep a vivid and controversial discourse about AI. It seems that AI systems will have an increasingly greater impact on our future lives and it is important that experts in this field contribute to the discussion in different ways. One is to acquire attention for the topic and illustrate serious importance and the other is to explain and provoke an opinion-building debate. However, the data provided is not representative, therefore I can not draw general conclusions from it.

The approach was supposed to observe what impact experts have who contribute to the public discourse. In my opinion, this approach can lead to further research but it would be important to have more reactions that can be evaluated. This would lead to possibly representative results and could avoid a speculative analysis where certain observations can not be mapped clearly to a causal explanation. However, it should be considered that readers who comment on an article, most likely do that because the article exposed them to certain emotions which lead to the reaction. This might result in a more extreme or emotional picture that is drawn from analyzing comment sections. After all, I believe that looking at qualities of the reactions to expert contributions can give an insight on how people perceive and understand AI.

References

Alsheibani, Sulaiman, Yen Cheung & Chris Messom. 2018. Artificial intelligence adoption: AI-readiness at firm-level. In *Pacis*, 37.

Bendel, Oliver & Georg Etscheit. 2017. Lassen wir die Finger davon. *die Zeit (print) and (Zeit-Campus)* 16(16/2017). 1–2. https://www.zeit.de/2017/16/kuenstliche-intelligenz-moral-maschinenethik-interview?.

Der Wissenschaftler Prof. Dr. Oliver Bendel. N.d. https://www.oliverbendel.net/index.html.

di Lorenzo, Giovanni. 2021. Wofür stehen wir? *die Zeit (print)* 9(9/2021). 1. https://www.zeit.de/2021/09/pressefreiheit-journalismus-gesellschaft-spaltung-politik.

Goedert, Bettina. N.d. *Readership data DIE ZEIT*. https://www.iqm.de/fileadmin/user_upload/Sonstige/International/Downloads/200313_ZEIT__Rate-Card_2020.pdf.

Precht, Richard David. 2020. Maschinen können keine Moral. *die Zeit(print)* 26(26/2020). 1–2. https://www.zeit.de/2020/26/kuenstliche-intelligenz-roboter-moral-gefahr-ethik?.

Prof. Dr. Richard David Precht. N.d. https://www.hfm-berlin.de/index.php?id=296&tx_persons_persons[person]=195&tx_persons_persons[action]=show&tx_persons_persons[controller]=Person.

Wolfangel, Eva. 2021. Wie viel Ethik verträgt Google? *Zeit-Online.* 1–3. https://www.zeit.de/digital/2021-02/google-ethik-timnit-gebru-technologie-forschung.

Zühlsdorff, Henning. N.d. *Richard David Precht zum Honorarprofessor ernannt.* https://www.leuphana.de/news/meldungen/titelstories/honorarprofessur-precht.html.

Chapter 5

Discourse about AI on Reddit

Richard Matschke & Till Nicke

In this paper we focused on public discourse about AI on the social media site Reddit. Our research questions were, what subreddits talk about AI, which ones are the most active in regards to posting activity and what are some frequent co-occurring keywords. For this we conducted a quantitative analysis, collecting posts containing the keyword AI for the whole year of 2020. Our analysis has shown that discourse regarding AI happens in a large variety of topics and domains. A list of the ten largest subreddits was distilled and analyzed further. The most apparent domain we found in that list was ”gaming” with four out of ten subreddits showing a connection to computer games. Other areas in the top ten ranking included, decision making tools, job searches, new research findings, applications and industry. Adding to that three subreddits were synthesized into word clouds for better visualization of co-occurring keywords. Moreover the overview given is aimed at encouraging further research in this area.

Keywords: AI | Reddit | Social Media | Pushshift | Word Cloud

1 Introduction

The internet provides a variety of ways to engage in discourse in a public manner. One particularly interesting online platform in that respect is the social media site Reddit, where registered users can post questions or other types of content such as photos, links to other websites and short blog posts. All this content is organized by subject and is grouped together in so-called 'subreddits', which users can subscribe to, in order to create their own personal News Feed.

What makes Reddit stand out from other large social media sites is its relatively high level of anonymity and that accessing the posts is possible without the need of creating an account. This leads to users with an account enjoying almost the same anonymity as users without an account. Another factor is that the grouping of posts into different topics via subreddits, make it more accessible for scientists as opposed to other more individualistic social media sites, like Facebook, Twitter etc.. These advantages and the large amount of publicly available data that is being generated make Reddit especially useful for researchers from various fields.

In this paper we built on an existing research project, specifically created for accessing Reddit data called the Pushshift dataset Baumgartner et al. 2020. The Reddit Pushshift dataset is a collection of social media data, made available with the purposes of facilitating researchers work. To give a better illustration, in the time period between June 2005 and April 2019, Reddit had no less than 650 million posts, with more than 5,6 billion comments posted in over 2,8 million subreddits, all stored and available in the Pushshift dataset Baumgartner et al. 2020. With the help of the Pushshift API, made available by the Pushshift dataset Research project we established the objective to address the following research questions:

- What subreddits refer to AI?
- Which subreddits received the most posts in regards to AI?
- What are some key terms that are used often together with the word AI?

2 Methodology

For the purpose of this research paper, we decided to collect Reddit data for the whole year of 2020. We used the Pushshift dataset API and set ourselves up to extract all submissions with the keyword "AI" in them. A script was written, that collected the title, the content of the post, usually referred to as the 'selftext' and the number of keyword occurrences for each post together with the subreddit, in which the post appeared. Table 1 gives an overview of the search criteria for the data collection process. All this information was then stored in yaml files. Due to traffic limitations of the API, the data had to be collected in batches and the script was refined to extract all posts, by their submission time on Reddit. The script had to loop through every hour of all 365 days of the year, since otherwise queries would have surpassed the API limit of 5500. In the yaml files every day has its own entry, such that every day could be analyzed individually. The days however were then grouped together in twelve yaml files, one for each month of the year 2020. The whole collection process took us approximately three Weeks, due to

Table 1: Search Criteria

Search string:	AI
Timeframe:	Jan 2020 - Dec 2020
Information collected:	’title’, ’selftext’, ‘keyword occurences’ ,’subreddit’

the large amounts of data, the API limit of 5500 and limited access to a stable internet connection. Another issue we encountered during the data collection phase was that the Pushshift server did not respond for longer periods of time and sometimes yielded less results than expected, e.g. one hour consisting of ca. 2000 entries and the next hour only having ca. 30 entries. Another inconsistency in the dataset can be seen, when looking at the monthly distribution of posting activity in Figure 1, as the number of posts in December is significantly lower compared to the other months. We tried to recollect the whole month of December twice, unfortunately with no difference in the obtained data. We therefore decided to continue with the amount of data we had collected up to that point. The scripts including all the data we collected is publicly available on the github repository: https://github.com/rmatschke/AI_public_discourse_on_reddit.

Once the data was collected, some processing steps were needed. The first script called PlottingResults.ipynb was written to determine how many subreddits mentioned AI in at least one of their posts in general and also to give a ranking of which subreddits had the most posts in regards to AI. Additionally we plotted overall posting activity for the year of 2020 (see figure 1). Finally another script called wordcloud_Generator.ipynb was written to analyze what other words co-occur most frequently with the term AI. It consists of a python function that takes a subreddit name as an argument, which then creates a large string out of all titles from all the posts we collected for this given subreddit and returns a word cloud for this subreddit as an output. Word clouds have a variety of advantages as data visualization tools and have also gained the attention of education researchers Heimerl et al. 2014. Adding to that the authors DePaolo & Wilkinson 2014 emphasize the usefulness of word clouds as graphical knowledge representations, with the benefits of not overloading the viewer with information but rather giving a quick intuition about a given text. This encouraged us to use word clouds to visualize the large amounts of data and to synthesize our findings that way.

While analyzing the data, we found 37337 different subreddits mentioning the keyword AI in at least one of its posts in the year 2020. However this number is to be seen with caution, as we did not clean the dataset. Some languages use AI not as an abbreviation for artificial intelligence, e.g. french, where it is a form

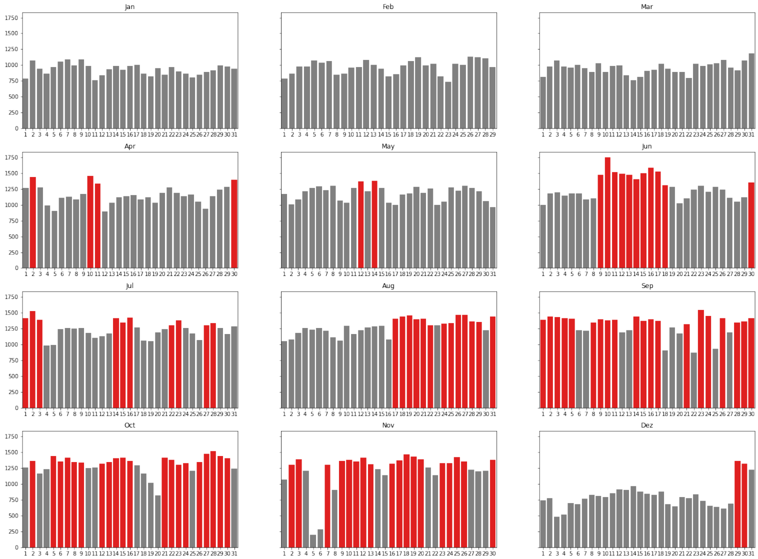


Figure 1: Distribution of the number of posts for all Months of the year 2020. Days with a post count above 1300 are marked in red

of the verb to have in the first person singular ”j’ai”. We tried to clean the data from these outliers, by using the python module langdetect. However by doing so, we would have lost more English entries, than non English ones. Therefore the decision was made to keep the raw data as it is and analyze it further in its uncleaned form.

3 Results

We found that a large portion of subreddits contained only one to about a handful of posts in a given month, on the other end of the spectrum were the subreddits who had well over a thousand posts over a one year period. Our analysis has shown the following ordered list of subreddits as being the most active ones in regards to AI, i.e. according to the number of posts containing the keyword AI: r/makeDecision, r/HeroesBattle, r/AIDungeon, r/ArtificialIntelligence, r/SteamGameSwap, r/indiegameswap, r/artificial, r/jobbit, r/MachineLearning, r/eu4.

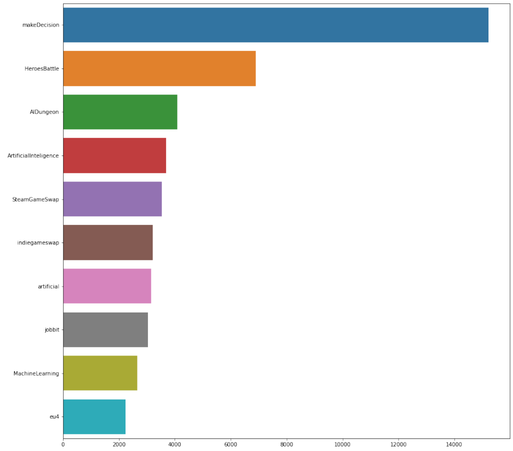


Figure 2: Top ten ranking of subreddits by the amount of posts containing the keyword AI

Figure 2 shows the ranking of this top ten list, with r/makeDecision being the most active and having 15225 posts and eu4 being on the tenth place with 2235 posts. Two additional rankings, namely a top 20 and a top 50 list can be found in the github repository. Looking at the top ten ranking in general we can see that discourse on Reddit is not limited to one specific area, but rather branches out into a large variety of fields.

A surprising finding for us was that the subreddits r/makeDecisions and r/HeroesBattle had the most posts regarding AI out of all the other 37337 subreddits. This was unexpected to us since they have a very small number of members, namely below 100 at the point of writing, (r/makeDecisions has 20 and r/HeroesBattle has 52, retrieved on 29.03 March 2021). Adding to that almost all posts we found in these two subreddits relied solely on a tool called AI.decider, which refers back to a website that apparently should help users make decisions between two options. The website states about itself: “AI.decider is a intelligent decider that helps you make important life decisions, set goals and improve your productivity using modern machine learning tools” http://aidecider.com/ 2021. A recommendation might be therefore to exclude those two subreddits in future analyses, since they are linked to a cooperate product, in this case AI.decider and generate little to no discourse at all. This is also indicated by a unusual little amount of comments that these posts receive. Even though we did not include the comments in our data, this specific outlier caught our attention, so we looked up this subreddit and all the posts we found had at most one comment, always from the same user.

Another interesting finding was, that a large proportion of discourse about AI on Reddit is happening in the area of Gaming. Four out of the top ten subreddits can be categorized as having a connection to Computer games, mainly r/AIDungeon, r/SteamGameSwap, r/indiegameswap and r/eu4. To give an example, the r/AIDungeon is a subreddit specifically for people interested in the text adventure game AIDungeon, which is powered by an artificial intelligence, generating unlimited stories and text based adventures for users to take part in.

Furthermore we decided to also present three subreddits, namely r/ArtificialInteligence, r/artificial and r/jobbit and their word clouds in a little bit more detail, since they showcase an interesting variety of topics related to AI. For more information about keyword occurrences of the remaining word clouds of the top ten ranking list the reader is advised to visit the above mentioned github repository. In the three subreddits r/artificial, r/ArtificialInteligence and r/jobbit and their respective word clouds, figure 3, 4 and 5, we can see various overlapping in topics. For example the coronavirus pandemic which started to gain public attention in 2020 has left significant trails in two of the three subreddits, such that the word clouds from r/artificial and r/artificialInteligence contain the keyword ”Covid” and other pandemic related terms. Even though there are other terms which cooccur between the different threads, the word clouds give the impression that they focus on different aspects of AI. We find words connected to ”research” and ”new” in r/artificial which suggests that the posts there are more centered around new findings in AI (see figure 3). Conversely r/ArtificialInteligence shows words connected to applications and industry, as seen in figure 4, where as posts in r/jobbit are specifically centered around different job titles, for people seeking work in the field of AI (see figure 5).

4 Discussion

Although the results gave us some interesting insights, there are a few limitations needed to be mentioned regarding our research. Due to time constraints we were not able to preprocess and clean the data, so that the collected data might include duplicates, posts from other languages, posts that were misclassified by a spelling mistake and so on. An important phenomenon to mention is cross-referencing in Reddit. This is the reposting of an article into a different subreddit. As it is difficult to track who posted what in which thread, we did not

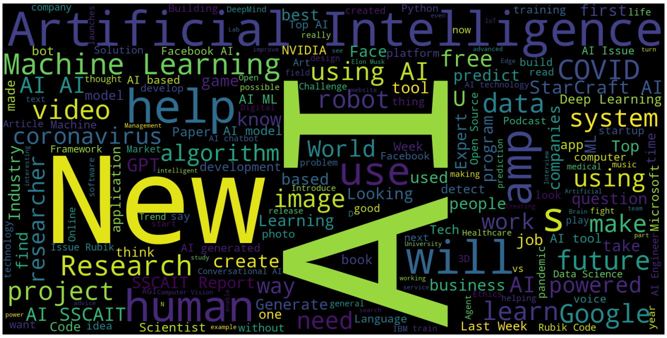


Figure 3: Word cloud generated from the post titles of the subreddit r/artificial

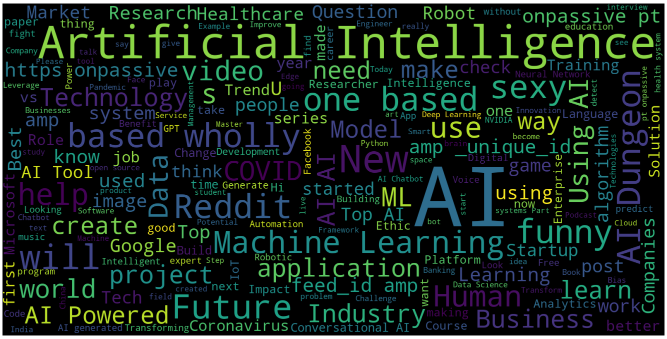


Figure 4: Word cloud generated from the post titles of the subreddit r/ArtificialInteligence

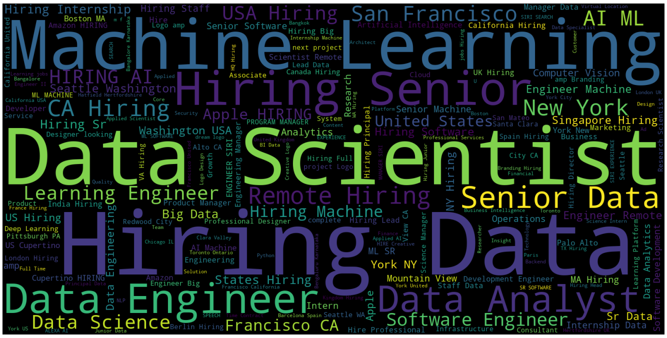


Figure 5: Word cloud generated from the post titles of the subreddit r/jobbit

clean the dataset according to the time stamps. Additionally there might be some missing data from the Pushshift dataset, leading to missing entries, this seems to be especially the case in the months of November and December, where some days show significantly lower number of posts than during the other months. Another important point to mention is that we do not know which of the posts we collected were created by bots rather than by actual humans. This can be seen in the cases of r/makeDecisions and r/Heroesbattle, where posts might have been computer generated to a large extent. Also due to constraints in time and computational power we were not able to collect data regarding comments, this is unfortunate since this type of data might have been especially useful to gather more insights about how discourse unfolded in regards to certain topics. Nonetheless we hope that our methodology including the data we collected and the results we obtained, can shed some light on the public discourse about AI taking place on Reddit.

5 Conclusion

The purpose of this work is to give researchers and people interested a good starting point for the investigation of discourse about AI on Reddit. Especially the ranking of the subreddits by submissions, the word cloud generating script and our insights in the limitation section might benefit future research in this field. We found that different subreddits are concerned with different usages of AI, e.g. gaming, new research, industry applications and job searches. What remains to say is that the field of AI is sparking discourse in a great variety of areas and domains and that there is large room for further research. We hope that this project can serve as a basis for future research work in this field and to have shown another interesting facet about AI in public discourse to the reader.

References

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire & Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 830–839.

DePaolo, Concetta A & Kelly Wilkinson. 2014. Get your head into the clouds: using word clouds for analyzing qualitative assessment data. *TechTrends* 58(3). 38–44.

Heimerl, Florian, Steffen Lohmann, Simon Lange & Thomas Ertl. 2014. Word cloud explorer: text analytics based on word clouds. In *2014 47th Hawaii International Conference on System Sciences*, 1833–1842.

http://aidecider.com/. 2021. *AI.Decider.* https : / / aidecider . wordpress . com/ (30 March, 2021).

Chapter 6

Twitter users' perception of AI in remote proctoring

Elen Le Foll & Lisa Titz

As a consequence of the COVID-19 pandemic, many educational institutions have turned to AI-based remote proctoring solutions to monitor students during online examinations. The deployment of such software has caused some concerns: reports of unfair treatment and accusations of racist and ableist systems have widely circulated in social media. Twitter has been at the forefront of these debates. The present study thus investigates how AI in online proctoring is portrayed by different stakeholders (e.g. companies, students, and representatives of educational institutions) in Twitter discourse. To this end, we compiled a small, highly specialised corpus of tweets (N = 467) which we manually tagged following a Qualitative Content Analysis (QCA) approach. The annotated data allows for the fine-grained analysis of the terminology used to describe AI in the context of remote proctoring in combination with the benefits, issues and emotions that Twitter users associate with such software. Our results suggest that general terms are preferred over more specialised ones. We also show that some terms are used inconsistently by the general public and sometimes even misleadingly by online proctoring companies themselves.

Keywords: Remote proctoring | E-proctoring | Online examinations | Anti-cheating software | Twitter | Suspicious behaviour | Facial recognition

1 Introduction

Not only has the COVID-19 pandemic resulted in a sudden switch to online and hybrid forms of teaching and learning, it has also forced schools, universities, and other educational institutions to find new ways of conducting examinations that respect social distancing guidelines. One of the major challenges that educational institutions have faced in organising remote examinations concerns the academic integrity of online assessment forms (Milone et al. 2017). Thus, in this new context of online distance learning, many educational institutions, and in particular universities, have opted for remote, online exam proctoring tools (Harwell 2020b). As a result, 2020 has seen an almost exponential increase in the demand for online proctoring solutions and "anti-cheating" software (Markets 2021). These services are usually provided by for-profit companies (Kharbat & Abu Daabes 2021). In spite of their high costs and widespread concerns over privacy issues (Harwell 2020b), their attractiveness is undoubtedly due to their ease of implementation as such systems enable the near replication of in-person exam conditions. They provide solutions that eliminate the need for the physical presence of humans proctors. Traditionally, human proctors supervise large groups of students answering previously undisclosed questions, most commonly with pen and paper, in a limited amount of time and with access to no, or only very limited, additional tools or resources.

Remote proctoring providers provide three types of remote proctoring solutions, all three of which make it possible to implement examination formats that are very similar or even identical to in-person assessments:

1. Live proctoring, which involves human proctors monitoring examinees using the video and audio streams from students' computers, thus simply removing the location constraint from traditional in-person examinations;

2. Fully automated proctoring, which involves AI-systems trained to identify fraud by recognising and classifying examinees' identity and activities as captured in real-time by the students' webcams and microphones; and

3. AI-supported live proctoring, which combines the two methods by having algorithms flagging and recording suspicious behaviour, which human proctors are to subsequently check to make an informed decision as to whether a student has indeed cheated, attempted to cheat, or whether the AI triggered a false alarm (c.f. LeewayHertz 2020, Singh 2021, Swauger 2020a,b).

In the second and third types of online, remote proctoring software, artificial intelligence systems are deployed for voice and facial recognition (e.g. to check that one and the same person is taking the test and that no one is assisting them), pattern recognition (e.g. to detect (potentially) suspicious head, eye, and computer mouse movements), speech-to-text (e.g. to monitor what is said) and object recognition (e.g. to ensure that test-takers are only using authorised additional tools and resources). Fully automated and AI-supported service providers and their proponents (e.g. Scott 2019) argue that the use of AI in remote proctoring improves the detection rate of dishonest behaviours by flagging suspicious patterns that humans may not necessarily be able to identify. Further, it has been hailed as a solution for making online examinations "credible and secure" (Scott 2019) since actions, such as excluding a dishonest student from an exam, can be taken immediately.

However, as the use of online proctoring tools has boomed, so has criticism of this new technology grown. Ever since the first wave of lockdowns in 2020, a number of grievances by students who claim to have been falsely accused of cheating by such software have gone viral on social media, spurring wider public interest in the matter (Harwell 2020a). One prominent example was college student and TikTok user @_.daynuh._who posted a tearful video on the popular video-based social media platform TikTok in which she reported:

> [...] my professor is giving me a zero, because the Review+[1] said I was talking when I was just, like, re-reading the question so I could better understand it (as cited in Harris 2020).

Within days, the video had reached more than three million views with many calling the AI "ableist" (Harris 2020) as it appears to unfairly disadvantage students with learning disabilities such as dyslexia. There have also been reports of racial discrimination with AI proctoring systems seemingly unable to detect the facial movements of people of colour (Swauger 2020b, Feathers 2021). A tweet by Alivardi Khan was cited in news sites from across the world and sparked outrage beyond the student community.

> The @ExamSoft software can't 'recognize' me due to 'poor lighting' even though I'm sitting in a well-lit room. Starting to think it has nothing to do with lighting. Pretty sure we all predicted their facial recognition software wouldn't work for people of color. @DiplomaPriv4All (as cited in Chin 2021).

In light of the growth and demand for (partially) automated, and therefore cheaper, online proctoring AI-based solutions and as a result of the reactions to these developments on social media, this study investigates the roles of AI in online proctoring as perceived by Twitter users. Twitter was chosen because it is the mouthpiece of a cross-section of different remote proctoring stakeholders. First,

[1] One of the products offered by the remote proctoring firm ProctorU, which advertises it as "The Economical Solution To DETECT Proxy Testing, Cheating & Content Theft [...] for low-to-mid-stakes quizzes and exams" (ProctorU 2021).

many online proctoring companies actively use this social media platform to advertise their products. Second, it is the home of a vibrant educational technology (#edtech) community. Indeed, many of the recent controversies around remote proctoring have either originated on Twitter, or quickly migrated to the platform (as in the case of the discussion around @_.daynuh._'s experience). And, third, it is also popular with students, some of whom have rallied to publish their experiences of online proctoring on specially dedicated accounts such as @Procteario and @ProcterrorU. We seek to explore the language used to describe artificial intelligence systems in remote proctoring systems in the tweets of companies providing such services, representatives of educational institutions, and students who are proctored by or with such tools. Section 2 outlines our data collection, data wrangling and manual annotation processes. Our quantitative and qualitative results are presented in Section 3. We conclude with a discussion of our main findings and an outlook for future developments in this fast-evolving field.

2 Data and methods

In the following, we describe our data extraction and processing processes, as well as the cyclical development and evaluation of a custom tagset for the qualitative analysis of the harvested tweets.

2.1 Data extraction

As our research question focuses on the perceived roles, benefits and dangers of AI in remote proctoring as communicated by Twitter users, the first step consisted in automatically downloading potentially relevant tweets using sets of keywords. We used the R (R Core Team 2020) rtweet package (Kearney 2019) to formulate and send requests to Twitter's REST API. Given the fairly low numbers of tweets involved and the fact that all tweets were to be subsequently manually sorted and annotated, the API queries were optimised for maximal recall rather than precision. We thus opted for a large set of API queries consisting of combinations of two keywords: one from a list of keywords related to online proctoring (I.), and a second keyword from a list of words and phrases potentially referring to some form of AI (II.). For each API query, two keywords were joined by the Boolean operator AND to form one query: e.g. "ProctorU AND algorithm"[2]. With our final selection of 12 proctoring keywords (I.) and 30 potentially AI-related keywords (II.), a list of all 360 possible combinations of the

[2] All queries were case-insensitive

two keyword lists was computed and subsequently used to automatically query the Twitter API.

I. Proctoring keywords: proctoring, online exam, online exams, remote examination, remote examinations, ProctorU, Examus, Proctortrack, AIProctor, ExamMonitor, Proview, Honorlock

II. AI keywords: AI, artificial intelligence, algorithm, algorithms, scanning, scans, scan, scanned, detection, detecting, detected, detect, recognition, recognise, recognize, recognises, recognizes, recognising, recognizing, recognised, recognized, automated, automation, automate, automates, machine learning, flag, flags, flagged, flagging

Up to 5,000 tweets (including tweet replies) per query were extracted on a weekly basis over a period of a month from 4 January to 1 of February 2021. In practice, the upper limit of 5,000 was never reached. The weekly intervals were chosen because, at the time the data was mined, the Twitter REST API granted access to "recent tweets" from the past 6-9 days (Kearney 2019) for non-commercial research purposes. Duplicates which were scraped more than once were eliminated from the dataset. In addition, regular expressions were employed to eliminate near-duplicates (e.g. tweets which were identical in their wording except for slightly different distributions of white spaces, or links or hashtags placed in a different order), which are typically sent out by companies or bots. This process resulted in 375 tweets for manual sorting and tagging. Given the modest size of the dataset, we decided to also specifically target nine Twitter accounts known to frequently contribute to Twitter discourse on remote proctoring and AI, and which we identified by searching for relevant keywords on Twitter. Six of these target accounts belong to companies that market online proctoring services with AI (@AIProctor, @conductexam, @examity, @ProctorExam, @proctorio and @ProctorU), two are anonymous accounts of detractors of online proctoring that mostly retweet student experiences and complaints (@Procteario and @ProcterrorU), and the ninth target account (@Linkletter) is of a learning technology specialist from the University of British Columbia who has been at the forefront of Twitter discourses on the use of AI in remote proctoring ever since he revealed the inner workings of Proctorio to the wider public in late August 2020 (Chin 2020). Although all the tweets published between 1 September 2020 and 2 February 2021 by these accounts were originally retrieved, only those that included at least one of the keywords from the AI list (II) were subsequently retained. This resulted in 200 unique tweets. When the two datasets were merged and duplicates removed, a total of 467 tweets remained.

2.2 Qualitative Content Analysis

As the next step, we developed a tagset based on the ontology collaboratively established by the authors of this book and the specifics of our dataset. Following an iterative Qualitative Content Analysis (QCA) approach (Kuckartz & McWhertor 2014), we evaluated our tagset on a random sample of tweets before applying it to our full dataset.

2.2.1 Development of the tagset

To do so, we first examined a random subset of 100 tweets from our full dataset. Having identified their authors' roles, concerns, and the terminology frequently used to refer to AI in remote proctoring by Twitter users, we derived guiding questions to describe our tagset. On this basis, we subsequently developed tags and tag definitions in an iterative process. The resulting tagset is outlined in Table 1. It is subdivided in nine categories which are briefly outlined below.

The first category of tags "Tweet: Properties of the tweet" (Table 1, first section) was designed to filter out irrelevant or uninterpretable tweets. Additional tags of this category which are not included in the table are: Tweets assigned the tag T_unrelated, which were deemed to be unrelated to our topic, the tag T_duplicate, which excludes near-duplicate tweets, e.g. those that only differ in their combinations of hashtags, and the tag T_difflang, which indicates that the contribution is not (fully) written in English. Finally, T_unclear was used when the content was not clearly about AI-based online proctoring services.

The second category in the tagset is concerned with the Twitter users involved in the discourse and aims to identify the distribution of contributions and opinions among students, educators, and companies.

The third category, definition, captures the aspects of the artificial intelligent systems that are discussed and the specific terms used to describe the technology. Both characterise and guide the understanding of AI within the discourse. We distinguish between the terms: "AI", "machine learning", "deep learning", "algorithm" and the words derived from the verb "automate". Tagged aspects of AI are behavioural patterns, such as hand or head movements, or turning around (D_behaviour) and, as special cases, eye movement detection (D_eye) and face recognition (D_face). In addition, we tagged discourse concerned with the amount of data gathered (D_data) and room scans (D_scan).

The next category focuses on issues associated with AI-based remote proctoring. The tag I_against describes a general negative attitude towards online proctoring and is assigned whenever none of the other tags within this category are appropriate, or whenever several issues are mentioned within one tweet, and at least one argument is not covered by the more specific issue tags. An example of a specific issue tag is: software failure (I_failed), i.e. tweets about systems falsely detecting cheating or demanding a human review. For face recognition problems of people of colour, the specific tag I_racism is used. Similarly, issues revolving around the unfair treatment of people with disabilities are tagged as I_disability. I_treatment is assigned to tweets that more generally refer to students having to follow strict rules, e.g. focus their gaze on the screen, or have their exams interrupted by the system to carry out room scans.

In contrast, the benefits category is concerned with discourse highlighting the beneficial aspects of remote proctoring, whereby the use of the B_helpful tag follows the rules of I_against and thus depicts a general positive attitude. B_fairness is assigned to tweets that contend that online proctoring makes examination situations fairer by detecting cheaters.

To investigate which emotions are related to remote proctoring, the tags E_positive and E_negative capture expressions that do not feature one of the more specific emotions listed in Table 1. Further, we also analyse language use with a specific tag for sarcasm (L_sarcasm), as well as tags to indicate that the current need for online proctoring solutions is construed as a consequence of the COVID-19 pandemic (C_covid and C_covidtrigger). The final category is concerned with the question of whether AI-based online proctoring should (continue to) be used in the future. Among those, the tag F_research is assigned to demands of increased investments and/or further research in the technology.

2.2.2 Evaluation and application of the tagset

First, both authors independently tagged 20 tweets using an earlier version of the tagset. We discussed the outcome and agreed on the addition, exclusion and refined definitions of certain tags. A second set of tweets was then tagged by both authors to evaluate the inter-rater agreement on this refined tagset. In total, just over 10% of tweets (52 out of 467) were tagged for this purpose. The 218 tags assigned to these 52 tweets by the two authors independently of each other were subsequently compared. Inter-rater agreement was found to be very high (94.2%). As a result, the second author tagged the remaining 415 tweets following the tagset outlined in Table 1. In the context of this qualitative tagging process, out of the various data fields retrieved from the Twitter API, only the tweeted text including any emojis and hashtags, the user screen name and, whenever possible, any linked web addresses were considered. Whenever necessary, we assigned several tags to any one tweet.

3 Results

In total, 467 tweets were manually tagged using the tagset laid out in Section 2.2. Of those, 215 tweets were excluded because they were tagged as either T_unrelated (N = 158), T_duplicate (N = 32), T_difflang (N = 16) or T_unclear (N = 9). Consequently, the final dataset analysed in the following consists of 252 manually annotated tweets.

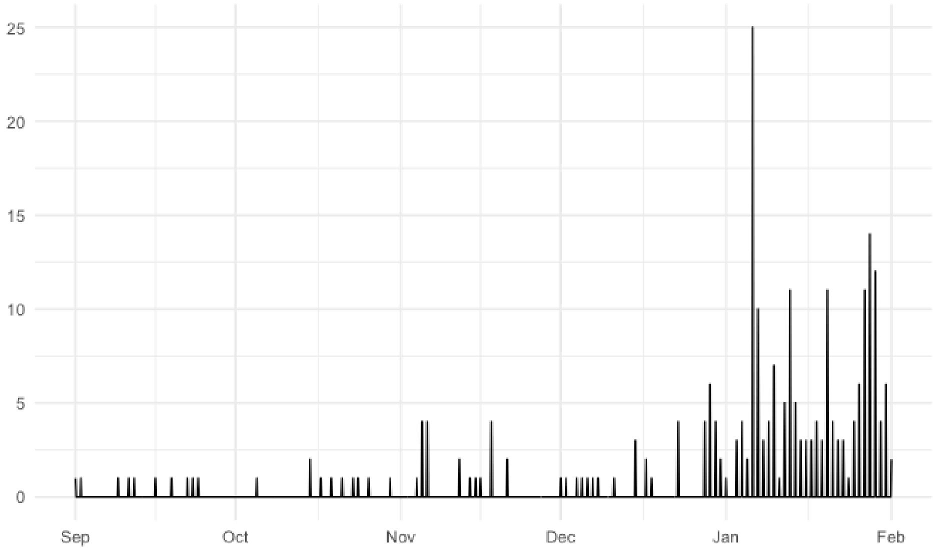


Figure 1: Frequency of tweets analysed (from 1 Sep 2020 to 2 Feb 2021, aggregated using three-hour intervals)

Figure 1 shows when these 252 tweets were published across the data collection period, which spanned from 1 September 2020 to 2 February 2021 (cf. Section 2.1). The outlier peak corresponds to the publication on 5 January 2021 of an article in The Verge on face detection issues with the online proctoring software ExamSoft, which was widely shared and discussed in the Twitter sphere (Chin 2021).

The distribution of the tags is depicted in Table 1 as counts and relative frequencies. The relative frequencies add up to more than 100 % because many tweets were assigned more than one tag. The distribution of identified user groups is listed in the first segment of the table. In total, 111 tweets were assigned to one of the three groups of users listed in Table 1, of which 65 tweets were identified as being published by a student, 28 tweets by a teacher or lecturer, and 18 tweets by a company providing AI-based online proctoring solutions. Second, we analysed which aspects of AI are most often discussed. As listed in the third segment of Table 1 tallying the terms used to describe AI, the most often assigned tags are D_ai (N = 63), D_face (N = 42), D_behaviour (N = 29), and D_scan (N = 25). In terms of definitions of the AI systems themselves, proctoring technology is most frequently referred to as involving "AI" (N = 63), whereas "machine learning" (N = 10) is used significantly less often and "deep learning" is not mentioned at all in our corpus. More general terms, such as "algorithm" (N = 14) and "automated/automation" (N = 16), were applied more often than the more specialist terms "deep learning" and "machine learning". Users seem not to associate the term ”machine learning" with their descriptions of tracking behavioural responses (pattern recognition), room scans (object recognition), or face detection and recognition. The only exception to this are two tweets quoting software promotions by Proctorio – an online proctoring service provider widely discussed throughout the tweets captured in the present dataset:

> ”Proctorio is the first and only proctoring solution that combines facial recognition technology and machine learning to eliminate any human error or bias" [...] (Tweet 447)

Only one tweet explicitly refers to biased algorithms by addressing the underlying mechanisms stating:

> [...] Aaaand yup it's a training-set problem ::angry face::[3] [...] (Tweet 117)[4]

Instead, the verb "flagging" is used particularly frequently by students to describe the fact that the software detects certain behavioural patterns, facial features, and/or object, and identifies them as inappropriate within the examination context. The following tweet exemplifies the use of the verb "flag" and the concerns of many students:

> i have to use online proctoring software for my exam next week and i'm so scared it'll flag me for something stupid ::crying:: (Tweet 169)

Moreover, face recognition is not only discussed as a software feature, but also to define the technology itself. The following two tweets serve to summarise the extensive Twitter-based debate on the extent to which the term "face recognition" can be used to define online proctoring software that rely on AI.

[3] Words bracketed by double colons represent emojis that were used in the original tweet.

[4] For data privacy reasons, tweets are only referred to by their position number in our dataset.

> A misconception about Proctorio is that we use facial recognition, but we never do. There is an important distinction between this & facial detection. Our choice to use facial detection means our software cannot be used to identify a test taker ever. https://bit.ly/3jVU0oX (Tweet 462)

The following tweet was published in response:

> Proctorio has now deleted the tweet where they said they did NOT use facial recognition. Yesterday they deleted the tweets where they said they DO use facial recognition. Can you see why people are confused? […] (Tweet 443)

Within the investigated discourse, the definitions of face recognition vs. face detection are fiercely debated. Hence, not only do users of online proctoring AI-based software not apply these terms according to consistent definitions, we also observe that one of the leading companies in the sector uses inconsistent terminology, seemingly in an attempt to defend its system.

Table 1: Tagset

Topic	Tags	N	Rel.F.
Tweet: Properties of the tweet			
Tweet covers topic, but cannot be matched to any other tag.	T_discourse	41	16.27%
Users: Are there common opinions among student/ teachers/ companies?			
To which group does the author belong?	U_company	18	7.14%
	U_student	65	25.79%
	U_teacher	28	11.11%
Definition: How is AI defined?			
“artificial intelligence” / “AI”	D_ai	63	25.00%
“algorithm”/ “algorithmic”	D_algorithm	14	5.56%
“automated”/ “automation”/ “automatic”	D_automation	16	6.35%
System detects behaviour.	D_behaviour	29	11.51%
Amount of data is gathered	D_data	16	6.35%
“deep learning”	D_dl	0	0
System detects eye movement.	D_eye	12	4.76%
System uses face recognition.	D_face	42	16.67%
“machine learning” / “ML”	D_ml	10	3.97%
System scans the room.	D_scan	25	9.92%
Issues: What issues come with remote proctoring?			
Generally against remote proctoring	I_against	34	13.49%
Generally biased system	I_biased	16	6.35%
Privacy issues	I_contraprivacy	20	7.94%
System is inappropriate for people with disabilities	I_disability	12	4.76%
Face recognition problems	I_faceprob	21	8.33%

Topic	Tags	N	Rel.F.
Software failed or is expected to fail	I_failed	44	17.46%
System makes more mistakes than humans.	I_moreerrors	0	0
Algorithm lacks transparency	I_opaque	13	5.16%
Racist software	I_racism	15	5.95%
How students are treated by system	I_treatment	23	9.13%
Benefits: What are beneficial aspects of remote proctoring?			
Allows access to education.	B_accessedu	5	1.98%
System supports fairness.	B_fairness	10	3.97%
System makes fewer mistakes than humans.	B_fewererrors	1	0.40%
Generally described as a helpful tool	B_helpful	11	4.37%
System respects privacy.	B_proprivacy	10	3.97%
Support education	B_supportedu	7	2.78%
Algorithm is transparent.	B_transparent	1	0.40%
Generally unbiased	B_unbiased	2	0.79%
Emotions: Which emotions are associated with remote proctoring?			
Positive emotions	E_hope	1	0.40%
	E_lessanxiety	1	0.40%
	E_positive	10	3.97%
Negative emotions	E_discomfort	1	0.40%
	E_fear	8	3.17%
	E_negative	52	20.63%
	E_worry	22	8.73%
Language: What kind of language is used?			
Sarcasm	L_sarcasm	9	3.57%
Covid: Is the discourse addressing the pandemic?			
Pandemic mentioned	C_covid	6	2.38%
Pandemic as trigger to use online proctoring	C_covidtrigger	0	0
Future: Should online proctoring be used in the future?			
Against	F_against	16	6.35%
For	F_for	6	2.38%
Further research demanded	F_research	10	3.97%

We sought to investigate whether users' contributions highlight positive or negative aspects of remote proctoring and thus to determine the extent to which AI is addressed in a positive or negative way in Twitter discourse. Our analysis of the tags pertaining to the issues associated with online proctoring reveals that 198 tweets mention (potential) drawbacks of the technology. By contrast, only 47 tags refer to beneficial aspects of online proctoring. Negative emotions are expressed almost seven times as often as positive ones. Unsurprisingly, whenever user role could be identified, negative emotive language was exclusively found in the tweets of students, as well as, perhaps less expectedly educators.

The following section investigates some of these negative aspects in greater detail. They are succinctly summarised in the following tweet:

> [I] had to take an exam with a proctoring software for the first time & that shit is so invasive & anxiety-provoking… when I take exams I'm not a robot… I have to move around, look around, & mouth words/talk to myself which are all things that could be flagged as cheating. dumb! (Tweet 375)

As made clear in the last part of the tweet above, students are concerned with the algorithm's categorisation of certain behavioural patterns or objects as suspicious (tag: I_failed). Additionally, students report that behaviour which interferes with face detection (e.g. touching their face) or eye movement detection (e.g. crying) is frequently and unfairly flagged as suspicious behaviour. Consequently, how students are treated within an exam situation (tag: I_treatment) is discussed very emotionally on Twitter. Concerns are frequently raised about the fact that these kinds of behaviours may result in being (temporarily) excluded from the assessment or in automatically failing the exam. Emotions repeatedly linked to these concerns include anxiety and anger towards the AI-based proctoring software (cf. Tweet 375).

The second part of Tweet 375 touches on the issue of privacy (tag: I_contraprivacy). The range and quantity of data gathered (tag: D_data) is the subject of much debate among opponents of the technology – with, here too, emotions of anxiety and anger frequently voiced, e.g.:

> I have to use proctorio for one [of] my online classes and I am honestly so terrified to use it. It basically scans your whole room and monitors what you are doing while taking an exam, which is a massive invasion of privacy and causes unnecessary stress. (Tweet 376)

Although the issue was raised by the press (cf. Section 1), racism was not a prominent issue in Twitter discussions on AI-based online proctoring during the surveyed period (N = 15 for I_racism). In the context of our analysis, the racism tag refers to reports or fears of AI systems' inability to correctly recognise the faces of people of colour. The topic was addressed in no uncertain terms by opponents of the technology, though without explicitly mentioning strong negative emotions, e.g.:

> In general, technology has a pattern of reinforcing structural oppression like racism and sexism. Now these same biases are showing up in test proctoring software that disproportionately hurts marginalized students. […] #edtech #equity #racism #sexism (Tweet 321)

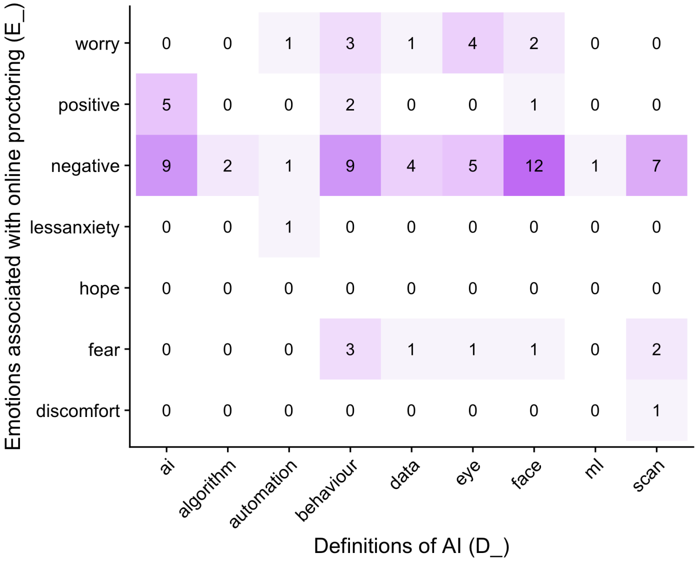


Figure 2: Heat map showing co-occurrences of the category Definitions of AI (D_) with the category Emotions associated with remote proctoring (E_)

To better visualise the co-occurrence of the tags that define AI with several other tag categories, heat maps were plotted. In the following, only the most frequently used definition tags, D_ai, D_face, D_behaviour and D_scan, are further explored.

Fig. 2 depicts how often certain definitions of AI are mentioned in combination with emotive language. The tags D_face, D_behaviour and D_scan are most predominantly used together with negative emotions (E_negative, E_worry and E_fear). In line with this observation, we also found that the most frequently used emoji was a crying face (N = 8).

Benefits of remote proctoring with AI-based solutions are almost exclusively discussed by company tweets (out of the tweets whose user role could be determined). Interestingly, as Fig. 3 shows, those descriptions mainly made use of superficial definitions of AI ("AI" and "algorithm") and variations of the word

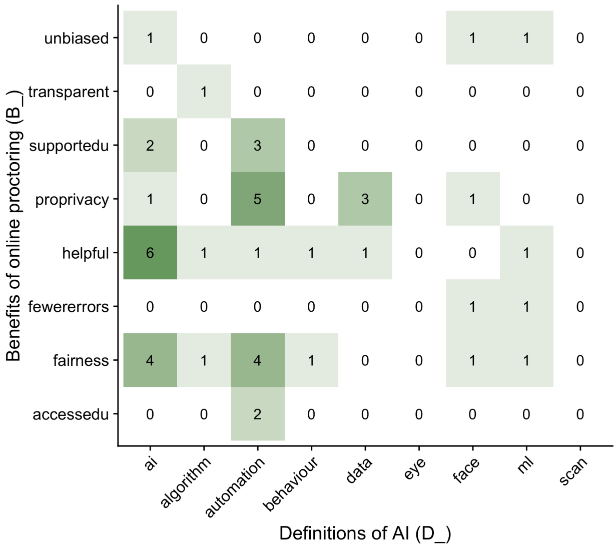


Figure 3: Heat map showing co-occurrences of the category Definitions of AI (D_) with the category Benefits of remote proctoring (B_)

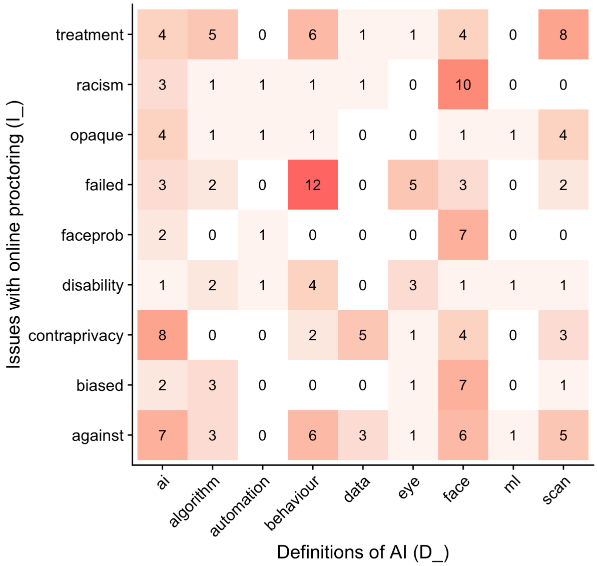


Figure 4: Heat map showing the co-occurrence of the category Definitions of AI (D_) with the category Issues with remote proctoring (I_)

"automate".

Further, we found that issues with online proctoring are exclusively discussed by students and educators (teachers and lecturers). As Fig. 4 visualises, the tags D_ai, D_face, D_behaviour and D_scan are applied frequently while users position themselves generally against online proctoring. Most of the problems and concerns with regard to face detection and recognition accused the AI of making racist and/or ableist decisions. Further, users report problems concerning pattern recognition of their behaviour, such as head or hand movements, and thereby discuss how they feel they are being treated while using such software. Additionally, examinees' improper treatment was frequently debated in the context of room scans.

4 Discussion

In our analysis, we saw that Twitter discourse is notable for its lack of clear definitions of the terminology used to describe AI in online proctoring software (Fig. 1, category definition). Contrary to our assumptions, companies do not lead the conversation around those terms and do not position themselves as experts in the field. This becomes especially clear in the Twitter-based debate on whether the software provider Proctorio applies "face detection" or "face recognition". Here, the company itself changed their statement several times without providing a clear definition or explanation of the terms or of the technology deployed in their marketed solutions. In the end, they claimed that they only use "face detection". To clarify (see Dwivedi 2018), face recognition builds on face detection as the algorithms involved in face recognition must first detect a face in order to proceed further. Face detection tracks the position of a human face in an image based on common human facial features. However, the face is not identified as belonging to a particular individual. According to Proctorio's web page, they offer "machine learning-enabled ID verification [which] automatically analyzes and captures images of test-takers and their ID's" (Proctorio 2021). In other words, they compare the facial features of individuals and thereby clearly apply face recognition algorithms. The information provided on their website thus contradicts the statements the company published on Twitter. Indeed the company tweets analysed as part of this study (e.g. Tweet 462) seem to have been purposefully written to, if not outright mislead the public, at least muddy the waters. As shown in Tweet 443, these corporate tweets led to genuine confusion about the correct use and definition of AI-related terminology among Twitter users.

While highlighting the beneficial aspects of their technology, companies only use superficial definitions of AI and make vague statements using terms such as "automation". This imprecise usage of terminology (Fig. 1, category definition) was also observed in the tweets of students. Students describe the roles of AI very well by using expressions like "flagging" (see Tweet 169) and "monitors what you are doing" (see Tweet 376). When they refer to the technology at all, they mostly use the general term "AI", and only rarely use more specific terms such as "machine learning", "object detection" or "face detection". Based on our findings, however, we cannot conclude whether the students in our dataset understand the underlying structures of AI-based proctoring systems but are not aware of more specific terms to describe them, or if they are choosing terms such as "flagging" as a means of self-identifying with the Twitter community of critics of the technology.

Our analysis revealed that tweets of students predominately discuss negative aspects of AI-based proctoring systems. Students describe being scared of using these systems (e.g. Tweets 375 and 376). This seems to be foremost caused by the systems' lack of transparency. For instance, users are not informed about which specific behavioural patterns may cause the system to flag suspicious behaviour; it is not clear to the students why AI-based proctoring systems seemingly proscribe many legitimate, routine movements. Similarly, students' anxiety is often triggered by the invasiveness of AI-based proctoring systems which cause a range of privacy concerns (e.g. Tweet 375). The tweets analysed in the present study make clear that examinees proctored with AI-based software commonly cannot reproduce decisions taken by the algorithms. This inevitably casts doubts as to the legitimacy of such decisions. Test-takers' anxiety is no doubt exacerbated by the fact that the underlying mechanisms of online proctoring technologies are not clearly communicated by the tech companies themselves.

5 Conclusion

Our study has confirmed that Twitter is a highly relevant sphere of public discourse in investigating the roles, benefits and drawbacks associated with AI in remote proctoring. Our analysis includes tweets from a range of stakeholders: including service providers, educators, representatives of educational institutions and, last but not least, students.

Our examination of perceptions of AI in online proctoring suggests that educational institutions would do well to consider the concerns that students frequently voice about unfair treatment and privacy violations before deploying online AI-based proctoring tools (cf., e.g. Kharbat & Abu Daabes 2021). Even in the context of a global pandemic, students' concerns should feed into discussions of possible alternative assessment forms (cf. Guangul et al. 2020). We find that companies marketing AI-based proctoring systems currently do not communicate how artificial intelligence systems are deployed in their software. As a result, their clients, both educational institutions as buyers of the services, as well as students and lecturers as the end-users are largely ill- or uninformed about what the AI-powered technology genuinely entails.

We suggest that greater transparency would likely benefit everyone. Companies could outline the steps they have undertaken to address known issues of unfair treatment and discrimination, and educational institutions could make informed decisions as to whether, when, and in what contexts relying on such systems may be appropriate (cf. Milone et al. 2017). Greater transparency would also empower test-takers to become more informed users of remote proctoring tools. Students would then be more likely to competently handle the technology without needing to face the fear and anxiety that the tools' opaque and non-reproducible mechanisms are currently provoking.

References

Chin, Monica. 2020. *An ed-tech specialist spoke out about remote testing software — and now he's being sued.* https://www.theverge.com/2020/10/22/21526792/proctorio-online-test-proctoring-lawsuit-universities-students-coronavirus (28 March, 2021).

Chin, Monica. 2021. *Examsoft's proctoring software has a face-detection problem.* https://www.theverge.com/2021/1/5/22215727/examsoft-online-exams-testing-facial-recognition-report (28 March, 2021).

Dwivedi, Divyansh. 2018. *Face detection For beginners.* https://towardsdatascience.com/face-detection-for-beginners-e58e8f21aad9 (30 March, 2021).

Feathers, Todd. 2021. Schools are abandoning invasive proctoring software after student backlash. *VICE.* https://www.vice.com/en/article/7k9ag4/schools-are-abandoning-invasive-proctoring-software-after-student-backlash (28 March, 2021).

Guangul, Fiseha M., Adeel H. Suhail, Muhammad I. Khalit & Basim A. Khidhir. 2020. Challenges of remote assessment in higher education in the context of COVID-19: a case study of Middle East College. *Educational Assessment, Evaluation and Accountability* 32(4). 519–535. DOI: 10.1007/s11092-020-09340-w. https://doi.org/10.1007/s11092-020-09340-w (28 March, 2021).

Harris, Margot. 2020. A student says test proctoring AI flagged her as cheating when she read a question out loud. Others say the software could have more dire consequences. *Insider.* https://www.insider.com/viral-tiktok-student-fails-exam-after-ai-software-flags-cheating-2020-10 (28 March, 2021).

Harwell, Drew. 2020a. Cheating-detection companies made millions during the pandemic. Now students are fighting back. *Washington Post.* https://www.washingtonpost.com/technology/2020/11/12/test-monitoring-student-revolt/ (19 November, 2020).

Harwell, Drew. 2020b. Mass school closures in the wake of the coronavirus are driving a new wave of student surveillance. *Washington Post.* https://www.washingtonpost.com/technology/2020/04/01/online-proctoring-college-exams-coronavirus/ (28 March, 2021).

Kearney, Michael W. 2019. Rtweet: collecting and analyzing Twitter data. *Journal of Open Source Software* 4(42). R package version 0.7.0, 1829. DOI: 10.21105/joss.01829. https://joss.theoj.org/papers/10.21105/joss.01829.

Kharbat, Faten F. & Ajayeb S. Abu Daabes. 2021. E-proctored exams during the COVID-19 pandemic: A close understanding. *Education and Information Technologies.* DOI: 10.1007/s10639-021-10458-7. http://link.springer.com/10.1007/s10639-021-10458-7 (28 March, 2021).

Kuckartz, Udo & Anne McWhertor. 2014. *Qualitative text analysis: a guide to methods, practice & using software.* Los Angeles: SAGE. 173 pp.

LeewayHertz. 2020. *Remote Proctoring Using Artificial Intelligence.* LeewayHertz - Software Development Company. https://www.leewayhertz.com/remote-proctoring-using-ai/ (28 March, 2021).

Markets, Research and. 2021. *Global $1.18 Billion Online Exam Proctoring Market to 2027 with COVID-19 Impact Analysis.* https://www.prnewswire.com/news-releases/global-1-18-billion-online-exam-proctoring-market-to-2027-with-covid-19-impact-analysis-301218806.html (28 March, 2021).

Milone, Anna S., Angela M. Cortese, Rebecca L. Balestrieri & Amy L. Pittenger. 2017. The impact of proctored online exams on the educational experience. *Currents in Pharmacy Teaching and Learning* 9(1). 108–114. DOI: 10.1016/j.cptl.2016.08.037. https://www.sciencedirect.com/science/article/pii/S1877129715200056 (28 March, 2021).

Proctorio. 2021. *ID Verification.* https://proctorio.com/platform/id-verification (30 March, 2021).

ProctorU. 2021. *Review+.* https://www.proctoru.com/services/review-plus (28 March, 2021).

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Scott, Aileen. 2019. *Artificial intelligence is making online proctoring safe and secure.* https://medium.com/@aileenscott604/artificial-intelligence-is-making-online-proctoring-safe-and-secure-9b03845602da (28 March, 2021).

Singh, Rajesh. 2021. Remote online exams made possible with the use of AI-based proctoring solutions. *MarylandReporter.com.* https://marylandreporter.com/2021/01/06/remote-online-exams-made-possible-with-the-use-of-ai-based-proctoring-solutions/ (28 March, 2021).

Swauger, Shea. 2020a. *Software That Monitors Students during Tests Perpetuates Inequality and Violates Their Privacy.* MIT Technology Review. https://www.technologyreview.com/2020/08/07/1006132/software-algorithms-proctoring-online-tests-ai-ethics/ (28 March, 2021).

Swauger, Shea. 2020b. What's worse than remote school? Remote test-taking with AI proctors. *NBC News.* https://www.nbcnews.com/think/opinion/remote-testing-monitored-ai-failing-students-forced-undergo-it-ncna1246769 (28 March, 2021).

Chapter 7

How do Twitter users receive AI-related field research conducted by police?

F. K. & F. P.

Between 2017 and 2018 biometrical face recognition was tested by the Berlin police at the train station Südkreuz. On Twitter, this caused a variety of reactions. We analysed 30 tweets with the hashtag Südkreuz in a qualitative manner and found that on this specific social network most reactions are negative. Broadly, there are three different groups of responses. One group criticises the technological aspects of the software such that the performance of the software would be too low and the statistical analysis flawed. A second group criticises surveillance aspects, missing privacy issues and concerns about civil rights coming with the introduction of such software. A third group is interested in possible adversarial attacks of the software systems in order to avoid identification. By that, this third group forms an intersection of the formerly described two groups.

Keywords: face recognition software | Twitter comments | police work | public opinion of AI | biometry

1 Introduction

Digitisation has pervaded more and more areas of life and seems to be ubiquitous. Thus, it is not surprising that security services like police authorities and intelligent services participate in digitisation processes hoping that new surveillance technology will make the life of citizens safer and more comfortable.

Between August 2017 and July 2018, a face recognition field study was conducted by the Berlin Federal Police. For one year three cameras at Bahnhof Südkreuz in Berlin were equipped with bio-metrical face recognition software. Hence, those cameras identified certain passengers with the help of artificial intelligence

(AI): 312 people volunteered as test subjects for this field study. Prior to the test, pictures of the participants were taken. Additionally, they they were issued Bluetooth transponders which they were told to carry with them, whenever they entering the train station. The AI software, based on neural networks, was trained on the pictures of the subjects. The signals emitted by the transponders the participants carried were recorded by receivers in the train station. Thus, a reference system was created which made it possible to check whether the test subjects were identified by the camera systems or not (Bundespolizeipräsidium 2018). The federal police claimed that with such technology the identification and arrest of wanted criminals would become easier and the train station a safer place. On Twitter, between 2017 and 2018 many users reacted to this field study in a lot of different ways.

In the following, different reactions will be analysed in a qualitative way. First, it will be explained how the respective tweets were collected in the first place and by which methods they were analysed later on. Afterwards, general findings that were revealed during the analysis will be introduced and it will be explained in what way the tweets were categorised. In a next step, the first of the categories used to classify the tweets will be discussed, namely those tweets concerned with scientific aspects of the test phase like technical and statistical details.It will furthermore be examined whether concerns about the face recognition systems are justified by taking into account the official test results issued by the Federal Police. Following, tweets will be discussed whose authors are concerned with non-technical aspects of the field study, e.g. privacy issues. Subsequently, a third group of tweets, namely those of Twitter users interested in adversarial attacks will be explored. In the end, the findings of our analysis and their implications are discussed and a brief outlook is given.

2 Methods

In order to analyse tweets regarding the face recognition trial, we focused on the hashtag 'Südkreuz'. On Twitter, hashtags are used to categorise tweets. This makes it easy for other users to follow the content they are interested in and to find the respective content. They can search for a certain hashtags and all tweets that include it will show up. Südkreuz is the name of the train station in Berlin where the field study took place. This hashtag was chosen to prevent any possible negative biases. It simply denotes the name of the station and by that can be seemed to represent a a rather neutral hashtag. Other, more biased hashtags used in this context were e.g. Verunsicherungsbahnhof (unsettling train station) or Überwachungsbahnhof (surveillance train station). Next, the tweets on Twitter using this hashtag in the years 2017 to 2019 were manually scanned. In 2017, the test phase started and lasted until August 2018. In the fall of 2018, the final report of the study was published and until early 2019 many twitter users referenced it. From all those tweets, 30 were selected and their content was analysed in greater detail. In the selection process it was made sure that the tweets were either related to AI and face recognition software or were sophisticated enough to allow for a greater analysis. Due to this, many tweets with the hashtag 'Südkreuz' could already be left out: A majority of users just criticised the police, politicians or the German state in general without any connection to the AI systems in use. Furthermore, tweets that only criticised surveillance measures or a general loss of data privacy without drawing any explicit connection to the field study were omitted. Other tweets proved to be too short or too meaningless to be analysed while again other users only posted pictures of the train station and the cameras without any comments except for the hashtags. Other tweets were not concerned with the immediate impacts of the AI systems but e.g. discussed security vulnerabilities of the transponders detected only after the the test had ended. Thus, such tweets were left out as well.

To do this, an ontology of tags was used to allow for a deeper analysis For one thing, it allowed to explore the (potentially political and scientific) backgrounds of the authors of the tweets (see 4 for more details). Thus, possible affiliations of the agents as well as their own AI experiences (if any) became apparent. For another, the attitudes and emotions authors potentially have towards the AI software were explored. Additionally, strengths and weaknesses as well as opportunities and threats that Twitter users attributed to the face recognition software were considered.

Finally, it was examined whether the Twitter that participated in the discussion users demanded actions or proposed certain operations that should take place due to the the introduction of the smart camera systems at the train station. Due to this analysis, we were able to put the tweets in broadly three different categories: Agents concerned with scientific aspects of the test phase, e.g. technical and statistical details, users concerned with non-technical aspects, e.g. privacy issues and a third group interested in adversarial attacks and thus combining aspects of the two former groups.

3 General findings

The analysis showed that in general in the Twitter sphere, there are predominantly negative responses regarding the face recognition study. Of course, this cannot be generalised to a broader public. Twitter opinions do not depict the dispositions and opinions of a general public towards this face recognition trial. Hence, no universally-valid and absolute statements can be made regarding whether the overall public is in favour or in opposition to the face recognition trial by the federal police.

In general, three categories in which the responses fall could be identified: Firstly, Twitter users criticise technical aspects of the AI systems used for face recognition. Especially a certain ineffectiveness mainly due to a high false positive rate was faulted. Most of those tweets were published after the release of intermediate findings of the BMI (Bundesministerium des Innern, für Bau und Heimat, Federal Ministry of the Interior, Building and Community)in December 2017 and especially after the release of the final report of the test phase in October 2018.

In a second group of responses, tweets regarding possible disadvantages of such smart systems can be pooled together. Topics of surveillance, privacy issues and possibly eroding civil rights are raised here. Users concerned with such topics usually tweeted about them right after the start of the project in July 2017 and continued to do so mostly throughout the one year project even though the majority of tweets was indeed published in the first months of the test phase.

There is an intersection between those otherwise quite distinct two groups of Twitter reactions, marking a third group of responses: Those users tweeted about possible adversarial attacks – ways to confound the camera systems by means of wearing certain clothing or by using special make-up and therefore preventing identification. Those tweets connect a certain technical knowledge about the workings of the AI systems in use with concerns about privacy and the wish to remain unidentified.

4 Backgrounds of users twittering about field study

According to their respective Twitter timelines and according to the tweets those people have issued on other, different topics, users who criticise the technical and statistical parts of the field study mainly have a background in computer science. Most of them are persons affiliated with information technology, programming and web developing. They use their twitter accounts as private persons, meaning that their accounts are not connected to any companies, political parties or organisations. On the other hand, other Twitter users write professionally for blogs like netzpolitik.org, a German blog offering information about digital rights and internet politics with high impact. All those Twitter users share rather left political views, and, while knowing a lot about the technological and mathematical background of the AI systems used for the field study by the police, they nevertheless respond critically to this field study and often to face recognition by AI systems in general.

On the other hand, according to their respective Twitter profiles and according to other tweets issued on different topics, those users worried about privacy and surveillance do not necessarily have a background in (computer) science. Instead they are laypeople interested in these topics, journalists and advocacy groups wanting to shine a light on issues of privacy and surveillance, or politicians of the left party (Die Linke) and the green party (Bündnis 90/Die Grünen) concerned with civil rights.

Finally, those Twitter users at the interface of the former groups who tweet about adversarial attacks against the face recognition systems are diverse. Some of those do have a scientific background in terms of computer science while others, according to their profiles, are interested in civil rights, privacy issues and politics but are no experts in the field of computer science and AI, like a former Green party member of the parliament of Hesse, Hans Christoph Boppel.

5 Tweets criticising statistical analysis and "intelligent" software

In the following, the tweets of the formerly mentioned groups will be discussed and analysed in more detail, starting with the group tweeting about technical and statistical aspects of the software. It will furthermore be checked whether their claims and assumptions towards the face recognition systems are justifiedl. For this, the publicly available final report of the test phase issued by the Berlin and Brandenburg Federal Police in 2018 will be taken into account in more detail (Bundespolizeipräsidium 2018).

Both after the presentation of intermediate results in December 2017 and especially after the publishing of a final report in October 2018, the BMI and other sources tweeted their findings: Six months after the introduction of the face recognition software at Bahnhof Südkreuz in Berlin, the BMI claimed in a tweet "vielversprechende erste Zwischenergebnisse“ (promising first intermediate findings) and asserted a 70% chance of identifying test subjects (Bundesministerium des Innern, für Bau und Heimat 2017). Again six months later, the BMI and other technical affiliated twitter users cited Horst Seehofer, federal minister of the Interior, saying "die Systeme haben sich in beeidruckender Weise bewährt, sodass eine breite Einführung möglich ist" (that the systems proved to be impressively successful so that a broad introduction is possible) (Bundesministerium des Innern, für Bau und Heimat 2018c, eGonvernmentComputing 2018). In follow-up tweets the BMI claimed a hit rate of their AI systems of more than 80% and an average false positive rate (FPR) of <0.1 % and thus even better results than those presented in the intermediate findings (Bundesministerium des Innern, für Bau und Heimat 2018b,a).

Following all those tweets, many Twitter users criticised the test phase and the statistical findings as "irreführende Augenwischerei" (misleading engagement in window-dressing) (Bahnhof Südkreuz 2018). Others wondered how such systems could even be called intelligent "angesichts des fatalen Scheiterns" (considering the disastrous failing) of them (Schwerd, Daniel 2018) and suspected the police of either having no knowledge about statistics or being outright bold about the results (Henning 2017).

In the Twitter community a false positive rate of <0.1% was not considered a good result, neither was the overall hit ratio of >80%, though both numbers were presented as good results in the final report. The users instead claimed that there are nowadays AI systems performing with a hit rate of 99% and that actually a hit ratio of only 80% makes such AI systems bad systems (Dachwitz, Ingo 2018). Additionally, Twitter were outraged by the missing communication of the quality measure of precision and by the low number of it when manually doing the maths oneself based on the numbers provided by the BMI (see e.g. (Jacoby, Julian 2018, Kryptomania 2019)).

Furthermore, they claimed that when actually introducing such face recognition software in real life, the number of false alarms and thus the number of innocent people being suspected of being wanted criminals would be very large. Thus, one Twitter user responded that one would then need to station an SEK there (SEK = Spezialeinsatzkommando, German special police forces) (Steiner 2018). Another user commented ironically: "Wenn man die Verbrechernadel im Heuhaufen nicht findet wirft man am besten noch tausende Unschuldsnadeln dazu, bevor man den Magneten holt" (If you're not able to find the criminal needle in the haystack the best thing to do is to throw thousands of innocent needles into it as well before you get the magnet) (ASK 2018) Other users fault the high financial costs that would come with the necessary police actions associated with the high number of false alarms (Kees 2018b).

In order to analyse whether the accusations of those Twitter users are actually justified, one has to take a look both at the intermediate numbers published by the BMI and at the numbers published in the final report six months later. It becomes apparent that those Twitter users who find fault with the proclaimed hit rate of >70% and a false positive rate of <1% for the intermediate findings, published six months after the test phase are started, were right. Assuming an average of 179,000 people passing through Bahnhof Südkreuz daily (Bundestag 2020), a hit ratio of 75%, a FPR of 0.5% and 312 test persons, the number of false alarm rises to 895 per day. It seems clear that it would not be feasible for the Federal Police to deal with such a large amount of false alarms.

Next, one can calculate the precision by using Bayes' theorem, meaning the probability of a person being a test subject when the AI systems say so (P(camera announces hit | ¬ test subject)). The probability for that drops to only 20%, proving the supposedly promising results claimed by the BMI as not quite fitting.

The same holds for the numbers given in the final report by the federal police in October 2018. When reading the report carefully, one notices that for the statistical analysis one assumed that only 1000 persons pass the cameras with the face recognition software in one hour (Bundespolizeipräsidium 2018). However, in 2019 Bahnhof Südkreuz was the 8th most frequented train station in Germany, with 179,000 passengers passing by each day, or 7548 people per hour (Bundestag 2020). It seems reasonable that this number did not change too much compared to numbers of passengers in 2017 and 2018 when the test took place. As to our knowledge, no further building projects or infrastructure projects were realised close by the train station in those years which might have led to an increase in passenger numbers. When calculating with those 179,000 people, the number of false alarms increases while the precision decreases drastically. (Precision ranging from 29.8% to 37.6% for the three camera systems; number of false alarms per day for the three camera systems ranging from 215 to 448 per day).

Again, those users claiming that the systems do not hold what they promise and are in no ways intelligent or well-performing, even though the numbers make a good impression at first sight, are right. One could indeed say that by assuming and reporting different numbers and leaving the fact aside that test subjects are rare readers are misled into false beliefs about the quality of the AI systems used.

6 Tweets criticising surveillance

As shown previously, the vast amount of false alarms would cost the police quite a sum of work and money. For example, the politician Saskia Esken tweeted about the face recognition test phase as well. She is now part of the dual leadership team of the SPD in the Bundestag and wrote in response to Horst Seehofer's remarks on the results of the experiment on Twitter: "Liebe @bpol_b, mal ehrlich: Die #Gesichtserkennung am #südkreuz ist nicht nur nutzlos, sonder sogar schädlich für Eure Arbeit und die Sicherheit. Und da haben wir über

Bürgerrechte noch nicht mal gesprochen." (Esken 2018) (literally: "Dear @bpol_b, in all honesty: The #facial recognition at the #südkreuz is not only useless, it is in fact harmful to your work and the security. And we didn't even talk about civil rights yet.") She is not the only person who expressed her anger on Twitter and who mentioned civil rights in the context of this experiment. For her and many others, the security and surveillance aspects are very important: The second group of tweets we identified mostly focused on non-technical aspects and issues like privacy, data protection and surveillance.

A vague fear about surveillance and doubt about the camera systems were noticeable in tweets like that of @montclairchen, even though they were coated with humour: "Oh Mann, müsste längst zu meiner Frisörin!! Die sitzt aber am Bahnhof #südkreuz! Werde wohl mit Mülltüte aufm Kopf zum Frisör müssen" (birgit 2017). She tweeted that she really needs to go to her hairdresser but since it is located at the Bahnhof Südkreuz, she will have to go with a bin bag on her head. (literally: "Oh no, I should have gone to the hairdresser long ago!! But she's located at Bahnhof Südkreuz! I guess have to go there while wearing a bin bag over my head"). This tweet thus does not make it clear what exactly the person is afraid of. While there is still a certain sense of humour in it it nevertheless becomes evident that the author of the tweet is still deeply uncomfortable having to cross Bahnhof Südkreuz and thus the smart face recognition cameras.

This tweet is symptomatic for an entire range of tweets. The authors cannot point out exactly what it is that scares them. It becomes evident that the mere thought of being able to be identified by smart security systems bothers them and makes them uncomfortable. They are afraid of their personal privacy being attacked by such AI systems and that it might be not valued high enough by official services. However, it still remains unclear by their tweets alone why exactly this is, where this fear comes from and what measures could be taken by the authors themselves or others (like e.g. safety authorities) to diminish their fears. They use Twitter to solely express their vague discomfort and indisposition and it becomes evident that they do not agree with the installation of such facial recognition software.

In contrast, other tweets were less fear based and calmly stated the facts. "Camera surveillance using facial recognition in public space is not in line with our constitution. It cannot rolled out without a federal law by the #Bundestag which will have to weigh all arguments, especially the results of #Südkreuz very carefully." (Leopold 2018), Nils Leopold tweeted in response to the vague fear of Big Brother in Berlin and thus of a surveillance state. In contrast to the before mentioned tweets, other users thus did not react in an emotional manner to the AI systems. Nevertheless, they were still sceptical toward it but based their doubts on facts. For many of them, it was important to mention that there are currently no laws in Germany allowing for, in their view, such a large invasion of one's privacy. Due to this they view the field study sceptical, consider it a breach of German law and are very sceptical about it. Those authors, however, remain bland and are content with simply tweeting about their concerns in a concise and straightforward way.

There are people who tweeted with a sarcastic spin on their scepticism. For example, Rudolf Lörcks tweeted the title of a Spiegel article which mimics the beginning of joke: "Treffen sich Orwell und Kafka am Bahnhof..." (Orwell and Kafka meet at the train station...) (Lörcks 2017). This alludes to Orwell's book "1984" and Kafka's novel "Der Process". Those writers metaphorically meet at the Bahnhof Südkreuz in this facial recognition experiment. "1984" is dystopian story about a surveillance state while "Der Process" deals with a common man being at the complete mercy of an unsettling, anonymous controlling power. Others also alluded to Orwell in tweeting "1984 hat angerufen, sie wollen ihren Innenminister zurück." (1984 just called, they want back their minister of the Interior) (Locke 2017) It is noteworthy that in this case, the critique was directed at the minister of the interior at the time, Thomas deMaizière, who was partly responsible for getting the facial recognition experiment underway. These kinds of allusions show people's worries about the state increasing its power with the technological progress bringing about evermore possibilities for surveillance and henceforth it turning into a surveillance state.

7 Tweets about possible adversarial attacks

Finally, there is the intersection between the two previously analysed groups. Those users commented on the possibility of adversarial attacks in order to confound the AI software in use. While some users offer some sort of scientific evidence how it would be possible to fool face recognition software, other tweets are less scientifically sound and rather wish to polarise or start a discussion about face recognition. Again, one notices that those Twitter users who in their tweets refer to scientific evidence have a more scientific AI background than those users who tweet in a more general manner.

To give examples, one user refers to a scientific paper about installing small LED lights in a baseball cap, that, when turned on, might prevent identification from AI face recognition software (Kees 2018a). Other users, in a jokingly manner, place great confidence in the use of gaffer tape: "Hält alles zusammen. Geht schnell wieder ab. Auch am Südkreuz gegen Gesichtserkennung nutzbar" (Holds everything together. Easily removed. Possible to use at Südkreuz against face recognition) (Boppel 2017). In general, many users write about make up that might prevent the systems from recognising persons. In an exemplary manner, yet another user tweets in a polarising way: ”Ich geh mich dann mal am südkreuz überwachen lassen. Irgendwelche Schminktips?” (I’ll go and have myself surveilled at Südkreuz. Any make-up tips?) (four seasons total hocus pocus 2017).

Still others bring into play the idea of wearing hats and glasses to fool the systems (Kees 2017). There is scientific evidence that wearing paper glasses imprinted with adversarial noise - a certain colour pattern especially designed to prevent recognition - fools face recognition systems to a high degree (Sharif et al. 2016). However, in order for it to work such noise needs to be created on an individual basis for each picture in the database of the software, such that the original picture the neural net was trained on would be required. Simply using traditional make-up, glasses or hats might not be enough to prevent face recognition by modern AI systems.

Even others bring up the idea of fooling the camera systems in an even more obvious way, as one user puts it: ”Kann mal jemand Fahndungsfotos ausdrucken und in die Kameras halten? Eine Art ’Denial of Service’ für die Überwacher” (Can someone just print mugshots and hold them into cameras? Some kind of ’denial of service’ for the controllers) (Weidmann 2017). While the idea of fooling AI systems in such a way is certainly charming and in a way appealing, of course, in reality this would not be possible. The systems were only trained on the pictures of test subjects, not other people, and additionally it would probably not suffice to hold those pictures into the cameras in order to bring down such an AI software.

8 Conclusion & outlook

As it was shown, the field study exploring facial recognition from AI software by the federal police at Bahnhof Südkreuz between 2017 and 2018 attracted mainly negative comments on Twitter. Mostly, three distinct groups of critique can be distinguished: one group of Twitter users focuses on the technical aspects of the AI software. Especially the hit rate of the software is in their focus which is often criticised as too low. Furthermore, the statistical evaluation of the findings is faulted at as Twitter users consider the false positive rate as too high. A second distinct group of comments deals with surveillance issues and privacy concerns. Users of this group fear that their privacy is at risk when being monitored by smart security cameras equipped with face recognition software. Furthermore, they are afraid that civil rights will become undermined when using such face recognition which is considered unlawful. Indeed, as the notion of face recognition software and their usage in public spaces has only evolved recently, there are no laws up to today regulating and managing its precise usage. Furthermore, it has not been made clear yet (e.g. by rulings of the Federal Constitutional Court) whether the usage of such systems violates any basic rights. Finally, a third group on Twitter mainly talks about possible adversarial attacks. Thus, they discuss certain ways to fool those smart camera systems. Among others, certain make-up is considered as well as different disguises, including sunglasses, hats with special LED-lights in them, and more profane things like the use of duct tape.

However, those findings cannot be generalised. No final conclusion can be drawn by examining at only a qualitative level a total number of 30 tweets posted in response to the test phase. Only because on Twitter a majority of users opposed facial recognition software this does not necessarily mean that in real life a majority of the (German) people does so well. More research would be needed on this matter to either confirm or oppose our findings.

As the digitisation of our society moves in an ever faster speed, questions regarding the usage of AI systems in order to improve safety and prevent crimes come up more and more. It is a pressing matter to find answers to them and by that answer the question in what kind of future society humankind wants to live and how one wants to deal with digitisation. Or, to use one final tweet regarding facial recognition software: Not the algorithms, nor the software is actually dangerous but rather the humans who use this technology in irresponsible ways (nd.Aktuell 2018).

References

ASK. 2018. *Wenn man die Verbrechernadel im Heuhaufen nicht findet wirft man am besten noch tausende Unschuldsnadeln dazu, bevor man den Magneten holt. #Gesichtserkennung #Südkreuz #Überwachung.* https://twitter.com/ASKtheUnknown/status/1074567752772452352. Twitter.

Bahnhof Südkreuz. 2018. *Die Erkennungsraten der Gesichtserkennung als "ausgesprochen leistungsstark" zu beschreiben wie Bundespolizeidirektion Berlin Striethörster, ist irreführende Augenwischerei. Hochgerechnet bekommen wir Fehlalarme im Sekundentakt.* https://twitter.com/_suedkreuz/status/975010121578438656. Twitter.

birgit. 2017. *Oh Mann, müsste längst zu meiner Frisörin!! Die sitzt aber am Bahnhof #südkreuz! Werde wohl mit Mülltüte aufm Kopf zum Frisör müssen.* https://twitter.com/montclairchen/status/907477998315728896. Twitter.

Boppel, Chris. 2017. *#Gaffatape | Hält alles zusammen | Geht schnell wieder ab | Auch am #Südkreuz gegen #Gesichtserkennung nutzbar.* https://twitter.com/ChrisBoppel/status/894510224693972992. Twitter.

Bundesministerium des Innern, für Bau und Heimat. 2017. *Nach vielversprechendem ersten Zwischenergebnis der Gesichtserkennung am #Südkreuz hat Minister #deMaizière erklärt: 'Bei 70 Prozent und mehr haben wir eine positive Erkennung der gesuchten Testpersonen - das ist sehr guter Wert.'* https://twitter.com/BMI_-Bund/status/941635030069202944. Twitter.

Bundesministerium des Innern, für Bau und Heimat. 2018a. *Danke für den Hinweis & ja ;) ... die Falschtrefferraten liegen durchschnittlich unter 0.1%. Dieser Wert lässt sich durch die Kombination verschiedener Systeme technisch auf bis zu 0,00018% und damit auf ein verschwindend geringes Maß reduzieren. 2/2.* https://twitter.com/BMI_Bund/status/1050416667468156930. Twitter.

Bundesministerium des Innern, für Bau und Heimat. 2018b. *Die zum Einsatz kommende Technik erleichtert es, Straftäter/innen ohne zusätzliche Polizeikontrolle zu erkennen und festzunehmen. Die durchschnittliche Trefferrate liegt bei dem besten getesteten System unter realistischen Testbedingungen bei über 80%, ... 1/2.* https://twitter.com/BMI_Bund/status/1050416667468156930. Twitter.

Bundesministerium des Innern, für Bau und Heimat. 2018c. *Testergebnisse zur Gesichtskennung am Bahnhof #Südkreuz veröffentlicht. #Seehofer: Systeme haben sich in beeindruckender Weise bewährt, sodass breite Einführung möglich ist. Können damit die Sicherheit für Bürgerinnen & Bürger verbessern.* https://twitter.com/BMI_Bund/status/1050416667468156930. Twitter.

Bundespolizeipräsidium, Potsdam. 2018. *Abschlussbericht des Bundespolizeipräsidiums zum Teilprojekt 1 "Biometrische Gesichtserkennung" am Bahnhof Berlin Südkreuz.* Tech. rep. Bundespolizei.

Bundestag. 2020. *Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Torsten Herbst, Frank Sitta, Dr. Christian Jung, weiterer Abgeordneter und der Fraktion der FDP – Drucksache 19/19475 – Verlässlichkeit des Schienenverkehrs an Knotenbahnhöfen.* Tech. rep. 20455. Deutscher Bundestag.

Dachwitz, Ingo. 2018. *Dr. Stefan Ullrich verknüpft seinen Vortrag über Erkennungsmechanismen von KI mit einem aktuellen politischen Thema: "Gute biometrische Erkennungssysteme haben eine Trefferquote 99%, schlechte wie die am #Südkreuz kommen nichtmal auf 80%." #öfit2018 #überwachungsbahnhof.* https://twitter.com/roofjoke/status/1052896256178176006. Twitter.

eGonvernmentComputing. 2018. *'Die Systeme haben sich in beeindruckender Weise bewährt, sodass eine breite Einführung möglich ist', kommentiert Innenminister Horst Seehofer zum Abschluss des Projekts '#Sicherheitsbahnhof #Berlin*

#Südkreuz'. https://twitter.com/egovcomde/status/1055054029066301441. Twitter.

Esken, Saskia. 2018. *Liebe @bpol_b, mal ehrlich: Die #Gesichtserkennung am #südkreuz ist nicht nur nutzlos, sonder sogar schädlich für Eure Arbeit und die Sicherheit. Und da haben wir über Bürgerrechte noch nicht mal gesprochen.* https://twitter.com/EskenSaskia/status/1052058866815516672. Twitter.

four seasons total hocus pocus. 2017. *Ich geh mich dann mal am #südkreuz überwachen lassen. Irgendwelche #Schminktips?* https://twitter.com/ihastwitta/status/899911290096943104. Twitter.

Henning. 2017. *Bei großflächig angedachten Tests eine Trefferquote von ~ 70% und eine Fehlerquote von "unter 1%" als gute Werte zu bezeichnen, zeugt entweder von Dreistigkeit oder völligem Unverständnis von Statistik. #Südkreuz.* https://twitter.com/ben_tinc/status/943913797030416384. Twitter.

Jacoby, Julian. 2018. *Wenn ich mit 300000 Personen rechne, was für Berlin sehr konservativ geschätzt ist, liegt die Precision bei 45%! Da können Sie beser einen Affen hinstellen, der durch Zufall die Personen errät. Denn der trifft wenigstens 50%! #Südkreuz #Gesichtserkennung.* https://twitter.com/JacobyJu/status/1051238072296325121. Twitter.

Kees, Benjamin J. 2017. *Fahndung mit Gesichtserkennung? 34% der Gesuchten werden Brillen und Mützen tragen. 82% auf ihr Smartphone starren... #Verunsicherungsbahnhof.* https://twitter.com/Algoropticon/status/892197883163922433. Twitter.

Kees, Benjamin J. 2018a. *Gesichtserkennung von @BMI_Bund, @DB_Bahn und Bundespolizei schon vor dem Launch kaputt. Studie zeigt: Mit unauffälligen LED-Lämpchen in der Mütze #Gesichtserkennung an der Nase herumführen und sich als jemand anderes erkennen lassen geht nun. Quelle: arxiv.org/abs/1803.04683.* https://twitter.com/Algoropticon/status/974335441301360640. Twitter.

Kees, Benjamin J. 2018b. *Hier noch mal fürs @BMI_Bund und @bpol_b zum nachrechnen, ob Gesichtskerkennung etwas bringt und wie viel Kosten durch Fehlalarme verursacht werden. #Verunsicherungsbahnhof.* https://twitter.com/Algoropticon/status/959011129853956096. Twitter.

Kryptomania. 2019. *"Von 14 Personen, die das System identifziert hat, sind somit nur 29 Prozent tatsächlich verdächtig" – der gute alte Satz von Bayes und die #Überwachung.* https://twitter.com/kryptomania84/status/1101732088888389633. Twitter.

Leopold, Nils. 2018. *Camera surveillance using facial recognition in public space is not in line with our constitution. It cannot be rolled out without a federal law by the #Bundestag which will have to weigh* *all arguments, especially the results of #Südkreuz very carefully.* https://twitter.com/NilsLeopold/status/1040516494843629568. Twitter.

Locke, John. 2017. *1984 hat angerufen, sie wollen ihren Innenminister zurück. #deMaizière #Videoüberwachung #Südkreuz netzpolitik.org/2017/minister-...* https://twitter.com/JohnLocke1689/status/900700123885424644. Twitter.

Lörcks, Rudolf. 2017. *#Videoüberwachung am #Südkreuz Treffen sich Orwell und Kafka am Bahnhof... spiegel.de/netzwelt/netzp... #gesichtserkennung #polizeistaat.* https://twitter.com/rfc2460/status/901457411222974464. Twitter.

nd.Aktuell. 2018. *#Algorithmen haben einen schlechten Ruf. Diese Programme bestimmen im Computerzeitalter das Leben der Menschen, ist ein gängiger Vorwurf. Dabei ist nicht die Technik gefährlich, sondern der Mensch, der sie verantwortungslos einsetzt, so @ennopark https://dasND.de/1098086.* https://twitter.com/ndaktuell/status/1031949055377723392. Twitter.

Schwerd, Daniel. 2018. *Verstehe nicht, wie man diese Videoüberwachungs-Technik "intelligent" nennen kann, angesichts des fatalen Scheiterns.* https://twitter.com/netnrd/status/1052172389147168768. Twitter.

Sharif, Mahmood, Sruti Bhagavatula, Michael K. Reiter & Lujo Bauer. 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (CCS '16), 1528–1540. Vienna, Austria: Association for Computing Machinery.

Steiner, Falk. 2018. *System war in Testphase 1 mit pol. Erkennung vergleichbarar Quali gefüttert. In Testphase 2 mit mehreren Fotos der Personen. Was bedeutet Fehlerakzeptanzrate von 0,34 bei einem Bahnhof mit Zehntausend Menschen? Man muss dann ein SEK dort stationieren. #Südkreuz #gesichtserkennung.* https://twitter.com/flueke/status/1050464835593543681. Twitter.

Weidmann, Stephan. 2017. *#Suedkreuz kann mal jemand Fahndungsfotos ausdrucken und in die Kameras halten? Eine Art 'Denial of Service' für die Überwacher.* https://twitter.com/stw_77/status/894096615824388097. Twitter.

Part II

AI and subject experts

Chapter 8

How do scientists explain AI to the broad public? — A TED talks analysis

Antonia Becker, Michelle Görlitz, Yannick Hardt & Clara Schier

In this paper, our goal is to investigate the portrayal of AI in TED Talks, as a means of public communication. For our purposes, we differentiate between two rather distinct groups of disciplines –- STEM (Science, Technology, Engineering and Mathematics) and the humanities. On the one hand, we hypothesize that STEM based speakers describe AI and related topics in a positive way. On the other hand, we theorize that speakers with a background in the humanities depict AI negatively. In order to test our hypothesis, we use quantitative and qualitative approaches. For both we focus on strengths, weaknesses, opportunities, threats, proposals and demands of action. We extracted these categories from an ontology that we previously agreed upon. The quantitative approach consists of a keyword-in-context search with Sketch Engine. For a more thorough picture we also conduct a qualitative analysis of the five most watched talks for both areas of research. Both approaches support our hypothesis that STEM based researchers seem to portray AI in a mostly positive way. Speakers from the humanities appear to focus on the negative aspects of AI. Due to some limitations in our approach, future research is needed in order to see if these results can be replicated.

Keywords: TED Talks | AI | Experts | STEM | Humanities

1 Introduction

In our analysis we want to focus on so-called Technology, Entertainment and Design Talks (short: TED Talks) which are freely accessible online. The first TED conference was held in 1984, and since 1990 the conferences have been taking place annually, hosting speakers from various backgrounds (Ted.com 2021a).

There is also the possibility to obtain a license to host an independently organized TED event following certain guidelines (see Ted.com 2021e). These events are called TEDx events and are typically organized by universities, cities or other organizations.

We chose this source for analyzing artificial intelligence (AI) in public discourse as according to Sugimoto et al. (2013) the TED website is the most popular conference and events website in the world. This indicates its prominence not only amongst scientists but also laypersons. We are interested in how scientists with expertise in AI explain its applications, benefits and disadvantages to the broad public, using TED Talks as an example of popular scientific communication. By now, TED conferences are no longer limited to its original topics (Technology, Entertainment and Design) but have been extended to almost all disciplines, aiming to present "Ideas Worth Spreading" (Ted.com 2021b). As of now, there are more than 3600 talks available to watch online (Ted.com 2021c). Traditionally, the talks are restricted to a maximum length of 18 minutes, however this is not strict and therefore leading to some talks exceeding this limit (Romanelli et al. 2014). Further, the talks are being transcribed and translated into over 100 languages, which allows sharing the videos with audiences across the world (Wingrove 2017). Presenters of TED Talks are mostly experts on a given topic and passionate speakers. However, only a minority of approximately 21% of all speakers have an academic background (Sugimoto et al. 2013). The average age of the presenters is 47, ranging from 12 to 94 years (Meier et al. 2020). A study by Sugimoto et al. (2013) further suggests that the typical online audience is young (age range 18-24 years) and well-educated.

In our paper we aim to test the hypothesis that scientists with a STEM (Science, Technology, Engineering and Mathematics) background tend to portray AI and its implications positively while scientists from the humanities rather focus on the negative consequences resulting from the maturation of AI research. We assume that to be the case because STEM based scientists are likely to work directly with AI, possibly even engineering it, which allows for a deep and thorough understanding of the technology, its application, possibilities and limitations. The focus on the technological aspects might lead to overlooking the societal implications. Scientists from the humanities domain on the other hand are possibly less likely to have such deep knowledge as they may have less hands-on experience with engineering AI systems. This allows for the possibility that humanities based scientists focus more on general scenarios, sometimes leaving out the technical possibilities of AI. We assume that their professional focus may increase the likelihood of negative attitudes, as the correct inference of limitations and possibilities cannot be guaranteed. Furthermore, we think the typical humanities scientist could be prone to focus more on potential societal harm and criticism due to the nature of their field.

2 Methodology

We filtered all the available TED Talks online for the specific topic of artificial intelligence. When we realized that searching for 'AI' as a keyword yields a selection of more than 600 talks, we decided to instead browse the website by topic. As a result of that we were able to access a list of talks about AI that were specifically labeled as belonging to the topic of AI (Ted.com 2021d). We took into consideration all 77 videos that were published before November 2020, when we started our analysis. We excluded all talks that:

1. Were not scientific (i.e. not held by academics but rather comedians or other entertainers)

2. Do not have an official transcript, which we needed to allow for the automated corpus analysis

3. Contained a dialog between two or more people

4. Were part of a TED-Ed lesson

Further, we decided to only include talks from scientists that could clearly be classified as having either a background in a STEM field or in the humanities. For our purposes, we defined scientists as people with an academic background, independently of their current professional affiliation. This left us with a total number of 45 talks to investigate. 36 of those talks had a STEM related background and nine a background in the humanities.

From these talks we extracted the transcripts in English and then created a corpus based on them in Sketch Engine (2021). Using the same tool, we further divided the corpus into two subcorpora, one containing the 36 STEM talks and the other one containing the nine humanities talks. Because of the different number of talks in the two subcorpora also the amount of words in each differs considerably, such that the STEM subcorpus contains 80,144 words and the humanities subcorpus contains 24,183 words. The keywords we used in the corpus analysis are tokens that we extracted from an agreed upon ontology that was developed as the basis of the book (see Chapter 2). Furthermore, we also used synonyms of the ontology keywords to ensure a broader corpus search. We classified the keywords as being related to either strengths and weaknesses or to opportunities and threats (SWOT) during the analysis. Furthermore, we selected a set of keywords referring to proposals and demands that the presenters put forward. Based on these keywords we performed a contextual analysis. On this foundation we evaluated whether there is a systematic bias towards predominantly positive or negative claims within each subcorpus.

Using Sketch Engine (2021) we then started the quantitative analysis focusing on the occurrences of certain terms or phrases within the respective subcorpora. For that matter we used the concordance tool to search for specified queries and their uses in the context of the TED Talks. We then continued to manually investigate the context in which a certain term is used and classified it as either a strength or weakness, an opportunity or threat or a demand of action or a proposal of action to get an overview if certain expressions tend to occur predominantly in optimistic or pessimistic contexts in each of the subcorpora. Whenever a keyword resulted in hits, but there were no references to either of our categories, we excluded it from our analysis. If a term appeared more than 25 times as the result of a query search in one subcorpus, we took a random sample of 25 occurrences to assess the sentiment of the results.

Additionally, we conducted a qualitative analysis using the most popular videos from our selection. We chose to examine a total of ten talks, selecting the five videos with the highest number of views from each discipline. A brief description of the respective talks is provided in Section 3. We chose to select the talks based on their popularity as we assume that these will have a stronger influence on the public perception of AI compared to the remaining talks. This implicitly assumes that the views of the talks can be interpreted as audience reach. Furthermore, we decided to judge popularity based on the number of views on Ted.com (2021d) as opposed to other platforms such as YouTube, as we observed the number of views to be noticeably higher and therefore more meaningful on the TED website. For the qualitative analysis we evaluated the selected videos and judged their valence based on:

1. Our subjective impressions with regard to the atmosphere created by the speaker
2. The attitude of the speaker towards AI
3. Expressed concerns and chances

The evaluation of each talk was conducted by two independent observers in order to reduce the bias in interpretation. Subsequently we discussed our findings and synthesized them. By adding the qualitative analysis we hope to gain further insight and extend the findings from the corpus-based ontology-driven quantitative analysis.

3 Summary of the five most popular talks

In this section we will shortly present the five most watched talks from the STEM and the humanities collection respectively. We will briefly summarize all ten talks here and go into detail about the analysis in the results section.

3.1 STEM

3.1.1 Robots that fly and cooperate (Kumar 2012)

The most popular STEM talk was held by Vijay Kumar, who is a roboticist. The talk mainly focuses on drones, their technology, their opportunities and the algorithms they use in order for the drones to orient themselves even without GPS signals. Kumar focuses on the positive aspects of his technology, for example that their drone robots could be first responders in dangerous situations. This talk is rather technical and focuses on a certain application of AI. Another important aspect of this talk might be that Kumar also names some challenges for drones but he always combines that with possible solutions.

3.1.2 How AI can save our humanity (Lee 2018)

In the talk 'How AI can save our humanity' Kai-Fu Lee, a computer scientist and investor, talks about the future of human labor as AI is going to eliminate the need for workers in routine jobs and simultaneously bring us wealth. With the use of deep learning, technology can make predictions with extremely high accuracy, which will liberate us from routine jobs in the future and enable us to spend our time with newly created jobs of compassion. According to Lee (2018), "AI will never replace us as loving beings".

3.1.3 Get ready for hybrid thinking (Kurzweil 2014)

This talk is held by Ray Kurzweil, a computer scientist, inventor and futurist, who elaborates in detail about the neocortex. Kurzweil presents mostly past findings and achievements about neuroscience and especially the neocortex and he further gives a very ambitious outlook on future developments with the help of AI such as hybrid thinking and nanobots. He shares a very positive foresight.

Kurzweil emphasizes the potential for transhumanism and foresees this as the future we will most likely experience.

3.1.4 What happens in your brain when you pay attention? (Ordikhani-Seyedlar 2017)

The AI and machine learning engineer Mehdi Ordikhani-Seyedlar presents the applications of brain-computer interfaces in his talk. Even though he does not mention AI directly, he presents examples for the use of brain-computer interfaces for neurological disorders or for assisting coma patients. These interfaces use for example machine-learning technologies to learn to interpret brain signals.

3.1.5 Robots with "soul" (Hoffman 2013)

This talk is held by Guy Hoffman, who is a roboticist and focuses on embodied cognition and intelligence in robots. Hoffman highlights the importance of human-robot-interaction and proposes that we should create robotic systems with a focus on imperfections as humans can relate better to such systems compared to the more "chess-like" (Hoffman 2013) systems that do not take risks. Through his research he found out that imperfect and improvisational behavior of robots seems to be preferred by humans. One of his main goals is to improve the image of robots, which shows his positive attitude towards robots and AI in general. This goal is motivated by his belief that robots will necessarily be a part of society in the near future.

3.2 Humanities

3.2.1 Connected but alone? (Turkle 2012)

In this talk, Sherry Turkle, a cultural analyst, talks about how our devices redefine our communication and how they have a rather bad influence on our connection to each other as well. She suggests that we as humans are likely to fall victim to our own comfort by mainly communication through our phones. Turkle elaborates on her critical remarks on technology and its societal implications.

3.2.2 What happens when our computers get smarter than we are? (Bostrom 2015)

Nick Bostrom, a well-known philosopher, takes a look at past and current human evolution. As research suggests, AI will most likely become as smart as human beings and even exceed us at a certain point within this century. He explains that the next big step in our evolution will be machine superintelligence and that this will likely be our last invention ever necessary. Bostrom discusses possible threats that most likely will occur when we reach machine superintelligence and warns about a lack of precautions (currently and in the future) that could be highly negative for humanity.

3.2.3 We're building a dystopia just to make people click on ads (Tufekci 2017)

In this talk, Zeynep Tufekci, a techno-sociologist and assistant professor, focuses mainly on big data collection and data security. Tufekci emphasizes how the constant gathering and selling of personal data might enable people in power to manipulate us through the use of intelligent technologies. She gives examples on how people can be targeted with specific content based on the data that is already available about them. The consequences can range from the radicalization of individuals to mass changes in voting behavior as seen in the US presidential election of 2016. In order for personalized advertisement algorithms to work properly, they require an enormous amount of data thus encouraging surveillance structures. In her opinion, one of the problems of AI and these big data architectures is that most humans no longer really understand how these complex algorithms work. Further, she stresses that the privacy protection is inadequate and ethics considerations are nearly nonexistent.

3.2.4 How deepfakes undermine truth and threaten democracy (Citron 2019)

Deepfakes are the topic of Danielle Citron's talk. She is a law professor and deepfake scholar, presenting the problematic nature of deepfakes. They are a result of machine-learning technology that manipulates/fabricates audio and video material indistinguishable from its original. She introduces the topic with an example of a woman who fell victim to a deepfake. A pornographic video was published, in which she could be seen. Citron speaks not only about deepfakes but also stresses the consequences victims have to face and how the judicial system does not cover any of these problems. In a world with realistic deepfakes, nobody can rely on audio or video proof anymore and the political impact is unimaginable. She informs, warns and demands societal and judicial change in her talk for the sake of the victims and democracy.

3.2.5 How do we learn to work with intelligent machines? (Beane 2018)

This talk is held by Matt Beane, an organisational ethnographer and assistant professor. It deals with the impact that the development of artificial intelligence has on people's working environment. Beane argues that optimizing productivity through implementing AI changes the way that people are trained for their jobs. One of his examples is a surgical trainee who is missing the opportunity to learn on the job since well-trained robots are taking over surgeries. He suggests that AI should be created to support humans without taking away the quality of their jobs.

4 Results

In the following sections we will present the results for the respective disciplines. In order to do so we focus on strengths and weaknesses, opportunities and threats, as well as proposals and demands of action as our key factors.

4.1 Strengths and Weaknesses

Artificial intelligence features a lot of strengths, but it also has some weaknesses. It is important for us to identify these strengths and weaknesses within our complete corpus. In this analysis we used nine keywords resulting in a total of 246 hits. Out of the 90 references that we considered as relevant for our purposes, 49 are attributable to strengths and 41 to weaknesses. In the following two subsections we will go into detail about the strengths and weaknesses separately for the respective disciplines.

4.1.1 STEM

In our STEM corpus we analyzed 168 out of the 246 hits in total. We were able to extract 38 strengths and eleven weaknesses from that. As three times as many strengths as weaknesses were identified, this seems to be an indication that STEM talkers ascribe more positive attributes to AI in their presentations. Nevertheless, weaknesses still can be found, which suggests that STEM talkers do not have as much of a positively biased opinion of AI as we hypothesized prior to our investigation. This trend can also be observed in our qualitative analysis of the top five STEM talks.

In the talk 'How AI can save our humanity' by Kai-Fu Lee, both strengths and weaknesses can be found (see Section 3.1.2). He explains that one weakness of AI is, that it cannot be creative, which in turn implies a certain job security, which is positive after all. As Lee's talk is mainly about deep learning he names data and pattern recognition e.g. from pictures as one strength of AI. Furthermore, he elaborates that AI can optimize various processes and advances speech recognition, machine translation, drones and many more applications. He concludes that AI will take over routine jobs while helping us create jobs that we enjoy doing. Meanwhile, Guy Hoffman elaborates on social interaction with robots (see Section 3.1.5). In his opinion embodied intelligence and imperfections are strengths of AI because they make a robot more relatable in interaction with a human. Additionally, he believes that robots interacting with humans and giving them a sense that the robot knows what it is doing is a strength of AI. The only weakness he sees is that some robots have a rather calculated approach when it comes to interaction, which then does not elicit the same positive emotions in humans. Another talk that nicely presents strengths of AI is 'Robots that fly and cooperate' by Vijay Kumar (see Section 3.1.1). In this rather technical talk about drones and their technology it becomes clear that robots have a lot of applications. To name one application, they could function as first responders in dangerous situations (e.g. biochemical leaks), as most of them are extremely robust and recover from nearly everything.

Through our qualitative analysis of the top five STEM talks we found that the speakers see a lot of potential and strengths in AI and speak mainly positive about it. The few weaknesses that are mentioned seem to be solvable or not that severe.

4.1.2 Humanities

Within our humanities corpus we analyzed 78 hits in total, where we were able to extract eleven strengths and 30 weaknesses. These numbers indicate that scientists with a background in the humanities tend to portray AI more negatively. This is also shown in the qualitative analysis of the top five humanities talks, as mainly weaknesses are portrayed.

Sherry Turkle brings forward many critical remarks on technology and its societal implications (see Section 3.2.1). In her opinion, social robots show our lost confidence in being there for each other, they create an illusion of companionship. She does not name any positive aspects of AI, while she portrays potential applications of AI as a weakness. Another insightful talk is 'What happens when our computers get smarter than we are' by Nick Bostrom (see Section 3.2.2). Although, he sees the strong optimization process as a strength, he still has far more risks in mind. One of the weaknesses he mentions is that AI in its current state is domain specific and therefore lacks generalization. Furthermore, he adds that the functionality of our cortex is still superior to machines and that there are various safety issues with superintelligent AI. To be more specific, AI systems could possibly see humanity as an obstacle to the optimization process for which we have created them. One other safety issue could be the problem of keeping the AI contained in a closed system. Techno-sociologist Zeynep Tufekci highlights a lot of negative aspects of AI, too (see Section 3.2.3). She does not mention any strengths, but an exhaustive list of weaknesses. In her opinion, AI is inexplicable as the algorithms are too complex to be fully understood. Moreover, she thinks that AI lacks ethics considerations as well as privacy protection and that it could be easily abused for manipulation.

As already assumed in the beginning, our impression that the humanities have a rather negative point of view of AI is confirmed through the qualitative analysis of the five most watched humanities talks.

4.2 Opportunities and Threats

Regarding opportunities and threats that are portrayed in the selected transcripts we found a large number of keywords which occurred in both of our subcorpora. We analyzed 823 hits in total, stemming from 33 keywords. Out of those hits we judged 230 as relevant. 105 occurrences were referring to an opportunity and 125 to a threat.

4.2.1 STEM

Starting off with the STEM corpus we analyzed 501 hits in total. We were able to extract 77 opportunities and 48 threats from them. These numbers indicate that STEM talkers have an overall positive attitude towards AI in their presentations. Yet, they do also seem to see the possible negative implications that certain uses of AI may inherit. In a more interpretative tone, these results seem to resemble an optimistic but reflected view on AI. The generally positive tendency towards AI is something that can also be seen when taking a look at the qualitative analysis of the five most watched STEM talks, which will now follow.

The first of the five STEM talks we want to cover here is ‘How AI can save our humanity’ by the computer scientist and investor Kai-Fu Lee (see Section 3.1.2). The opportunities he lists are on the one hand monetary but on the other hand also humanistic. In more detail that means that he proposes the idea that AI will increase the annual gross domestic product (GDP) and that the collection of huge amounts of data will enable new entrepreneurial opportunities. He explains this with the example of China and attributes that to its current laws in regards to data. On the humanistic side he points out that the loss of routine jobs can be liberating in general. Yet, he sees the risk that the transitional phase in the work sector comes with. Furthermore, he explains that the GDP change and the shift of workforce can enable new careers that focus more on a side of human labor that is fueled by compassion and creativity. Something that AI cannot compete with. He concludes that we should embrace AI were it is needed and that we should focus on rethinking what makes us human, i.e. in regards to our work ethic. The next talk covered here is ‘What happens in your brain when you pay attention?’ by Mehdi Ordikhani-Seyedlar, an AI and machine learning engineer (see Section 3.1.4). The scope of this talk is purely positive and is concerned with the use and application of brain-computer interfaces in different settings. He lists a few examples for that: the therapeutic use in attention deficit hyperactivity disorder (ADHD) therapy today and the possibility of “reading thoughts” (Ordikhani-Seyedlar 2017) of people who are unable to speak (e.g. coma patients or stroke victims). He emphasizes the analysis of brain patterns as a means of achieving that goal. Ray Kurzweil, a computer scientist, inventor and futurist presents the idea of transhumanistic development that can be enabled through AI in synthesis with our neocortex in his talk ‘Get ready for hybrid thinking’ (see Section 3.1.3). He describes that possibility as the “next big leap for humanity” (Kurzweil 2014). Again this is a talk that solely focuses on the positive sides of AI but on a still rather theoretical and futuristic level. Another talk covered here is ‘Robots with soul’ by Guy Hoffman, a roboticist (see Section 3.1.5). He stresses the importance and future relevance of human-robot-interaction and highlights different approaches for it, ultimately proposing his favored one of those followed by some explanations. His research suggests that imperfect, improvisational behavior of robotic agents seems to be favorable over more calculated approaches. One could conclude that he implies that AI should be less accurate in its predictions in order to make interaction with it more pleasant, positive and seemingly relatable. The last talk we are covering here is ‘Robots that fly and cooperate’ by Vijay Kumar, who is a roboticist as well (see Section 3.1.1). To be more specific, this talk is about the use and development of drones. The main opportunity that we could extract from his talk is that the technology that he presents can have a lot of positive use cases, when applied. In general this talk is very balanced in terms of positive aspects, challenges and possible solutions.

Overall, we found that the presentations of STEM speakers focus mostly on positive aspects, while not ignoring possible challenges. The results of both the quantitative and qualitative analysis mainly support our hypothesis.

4.2.2 Humanities

Within the Humanities corpus we analyzed a total of 322 hits where we found 28 references to opportunities and 77 references to threats that AI could bring about. As we can see in this ratio, scientists with a background in the humanities clearly tend to shed light on potential hazards rather than on the opportunities. We have found a similar trend in the most popular talks from the humanities as will be shown in the following examples.

In his talk, Nick Bostrom speaks about risks and possibilities of machine superintelligence (see Section 3.2.2). Even though he also mentions considerable advantages such as strong optimization, he talks far more about threats of superintelligent AI. More specifically, he makes the point that if a super-intelligent AI is developed, humans will experience a "loss of control" (Bostrom 2015) which poses a great danger in case we do not plan ahead. Similarly, Zeynep Tufekci highlights the hazards that the constant collection of big data can pose on our society in her talk (see Section 3.2.3). She emphasizes how the abuse of data by people in power can be a "threat to our freedom and dignity" (Tufekci 2017) by controlling what is shown to every individual on social media platforms and therefore manipulating us. As an example, she mentions how elections can be manipulated by targeting individuals with specifically adjusted posts that are hidden to the public. Further, she addresses the threat that authoritarian regimes such as China may use AI technologies like face detection in order to identify and arrest people. Law professor Danielle Citron focuses on the threats that deepfake technologies may pose to individuals (see Section 3.2.4). She worries how the availability of these technologies may be misused and lead to cybermob attacks. In her talk, she does not mention any opportunities of AI or deepfake technology in particular but argues how dangerous the potential of deepfakes is.

These findings from both the quantitative as well as the qualitative analysis support our hypothesis that in their talks, scientists from the humanities domain focus on rather negative aspects (i.e. threats) of artificial intelligence.

4.3 Proposals and Demands

When scientists talk about AI, most of them refer to the future as in new evolving technologies, exciting opportunities and potential threats. Resulting from these topics we expected them to talk about solutions for potential problems and ideas to use the upcoming chances. But surprisingly we only found 29 references to proposals or demands of action in our analysis of 45 talks. We defined proposals of action as a suggestion on how to deal with a future situation, while demands of action are according to our definition an urgent call for action in order to face menacing circumstances in an adequate manner.

4.3.1 STEM

In our STEM corpus, we analyzed 68 hits where we found eight references to proposals and five to demands of action. The ratio between both types of references is relatively balanced. That suggests that the scientists with a STEM background mention proposals and demands of action equally.

In the talk 'How AI can save our humanity' Kai-Fu Lee talks about the future of human labor as AI is going to eliminate the need for workers in routine jobs (see Section 3.1.2). Even though the development of the job market is a big threat to the majority of people, Kai-Fu Lee is very optimistic that we will adapt to the new situation. He suggests creating jobs of compassion, as a substitute to all the routine jobs that are going to vanish from the human job market. In his vision, humans will be able to focus on creative work and make "labors of love into careers" (Lee 2018). Turning from the professional to the private sector the roboticist Guy Hoffman talks about social interaction with robots (see Section 3.1.5). He proposes to implement imperfections into robots to make them more relatable.

During our qualitative analysis of the five most watched STEM talks we got the general impression that the speakers were mainly optimistic concerning the future. Whenever demands of action were formulated, they did not seem to be dramatic and the situation never seemed unsolvable.

4.3.2 Humanities

The ratio of references seen in the humanities corpus is quite the opposite to the balanced ratio of the STEM corpus. In our analysis, we assessed 79 hits in total, where we did not find any references to proposals of action and 16 references to demands of action. This outcome combined with the findings of our qualitative analysis showed a clear tendency towards negative portrays of AI and its consequences in talks held by scientists from the field of humanities.

The talk 'How deepfakes undermine truth and threaten democracy' is a good example of this (see Section 3.2.4). Danielle Citron addresses the highly problematic upcoming of deepfakes. Due to a legal vacuum, there is no persecution of publishers of such deepfakes. In order to deal with this situation adequately, Citron advises social media platforms to take action and ban harmful deepfakes by changing the terms and conditions. Furthermore, she proposes to rely on human judgement for categorizing deepfakes in harmful impersonations and art/satire. But of course, this is not all we need to do to prevent deepfake induced harm. Citron also demands a proactive international solution and cooperation from technology companies, lawmakers, law enforcement and the media all around the world. According to her, everyone needs to be educated about deepfakes, especially journalists and law enforcers. To summarize the impression that her talk left, she is arguing factual and very demanding, which seems legitimate regarding the topic of her talk. Another even more persuasive talk in regards to our hypothesis, which is among the five most watched humanities talks is 'We're building a dystopia just to make people click on ads' (see Section 3.2.3). Zeynep Tufekci's talk deals with a highly discussed problem of the modern days, big data collection and data security. By portraying online advertisements as persuasion architectures she tries to raise awareness of what actually happens every day. Algorithms use our data to make us buy products, but they do not stop there. By analyzing our online activity the algorithms learn about our interests and try to suggest other content we might like and even start to auto-play new media. In the process of doing so the algorithm suggests more and more extreme content, going from vegetarianism to veganism or from slightly right-oriented political content to extreme or even radicalizing content. Tufekci demands a change by mobilizing technology, creativity and politics, so that AI can work constrained by human values. The change she is demanding is a digital economy that does not sell our data and attention. The philosopher Nick Bostrom has a similar demanding attitude towards the further development of AI (see Section 3.2.2). In his talk he demands that we start planning ahead to avoid harm in the future, which is currently predestined in his opinion due to human overconfidence with regard to AI. Planning and starting a conversation is also what Sherry Turkle demands (see Section 3.2.1). Talking about the societal implications of technology the cultural analyst shares her critical view and discusses potential threats our society will face in the future. According to her, social restructuring is what we need in order to adapt properly.

What is interesting in humanities talks is that they often focus on negative aspects of AI, portray a dramatic future, demand change, but rarely propose a concrete plan. To conclude the results of our quantitative and qualitative analysis for proposals and demands of action in both subcorpora, we were able to identify a pattern for humanities-based talkers to address the negative potentials of AI, while scientists with a STEM background show a rather positive attitude. This outcome confirms our hypothesis.

5 Limitations

Even though our analyses provide a seemingly clear picture, the following limitations should be mentioned. Time resources were the biggest limiting factor, which led to a rather superficial quantitative analysis and a rather small selection of talks for both analyses. Due to the broad range of older and newer talks ranging from 2003 to 2020 in the quantitative analysis and from 2012 to 2019 in the qualitative analysis, we have to consider the differences in technological development in that time frame, which influenced the content of the talks.

5.1 Quantitative Analysis

We only included a small selection of talks for our analysis. Whenever we found more than 25 occurrences for one keyword in the corpus analysis, we only analyzed a random sample of 25 hits in order to restrict the workload to a manageable amount. By searching for keywords in all talks available on the TED website we might have found more references to strengths, weaknesses, opportunities, threats, proposals of action and demands of action related to the topic of AI. Due to the inequality regarding subcorpora size (STEM: 80,144 words, humanities 24,183 words), we restricted our analysis to intrasubcorpora comparisons. Furthermore we would like to mention the fact, that our analysis was heavily influenced by the choice of keywords and other keywords might have resulted in different outcomes.

5.2 Qualitative Analysis

Our selection of talks for the qualitative analysis was based on the assumption that the number of views correlates with popularity and therefore indicate audience reach. A possible confounding variable in this correlation might be the time span the talks were available online, which is for obvious reasons greater for older talks. In virtue of selecting the most viewed talks we also have to assume the existence of a recommendation algorithm on the website, that probably has a biased suggestion technique. On the basis of this premise we need to indicate that the most popular talks might also be the most polarizing ones, from which we deduced a certain bias towards polarizing talks in our selection.

Furthermore we identified three out of ten qualitatively analyzed talks, which do not mention AI or one of its most common applications (e.g. Machine Learning) directly. These three talks are all from the STEM category: 'What happens in you brain when you pay attention?' (see Section 3.1.4), 'Get ready for hybrid thinking' (see Section 3.1.3) and 'Robots that fly and cooperate' (see Section 3.1.1). This circumstance may have influenced our perceived general attitude of the speakers and their portrayal of AI.

6 Discussion

The aim of this paper was to investigate the hypothesis that scientists with a STEM background tend to portray AI and its implications positively while scientists from the humanities rather focus on the negative consequences resulting from the progress being made in AI research. This hypothesis is confirmed through our quantitative and qualitative analysis.

In general, we found that STEM based scientists indicate a more positive view on AI. This might be due to some speakers being employed in the private sector which could translate into a personal interest in the success of AI systems. More precisely, STEM speakers see a lot of potential and strengths in AI, but also mention potential negative implications as well. However, a solution is usually provided and when they demand something, it does not seem dramatic or unsolvable. Whereas it does appear inevitable for speakers from the humanities. Scientists with a background in humanities seem to have a rather negative view on AI, which confirms our hypothesis once again. Humanities based speakers see some strengths in AI, but all the more weaknesses. Furthermore, they focus on threats and tend to shed light on potential hazards instead of potential opportunities. The most striking results however were found for proposals and demands. Humanities speakers portray a dramatic situation and demand change but mostly do not propose a solution. Overall, our results demonstrate that scientists with a STEM background tend to portray AI and its implications positively while scientists from the humanities rather focus on the negative aspects of AI.

Limitations are an important part of our paper, as we did have a limited time frame and capacities and therefore we only concentrated on ten talks in depth for our qualitative analysis. The constraints that are mentioned in Section 5 do restrict our analyses, as they lower the external validity of our study and limit the generalization of our findings. Despite these restrictions our findings seem to be clear. Regardless, future research should try to replicate our results with a larger corpus and a more exhaustive qualitative analysis. In addition, a different corpus i.e. one that was not tagged as AI by the TED Talks website, but includes every TED Talk that is somehow AI related, might be appropriate. Furthermore, the keywords used for the quantitative analysis could be extended. However, our findings provide a good starting point for discussion and further research. Nevertheless, future investigations are necessary to validate our findings.

Lastly, it is important to note that although STEM and humanities portray AI rather differently, both fields play an important role in multidisciplinary AI research and subsequently in scientific communication.

References

Beane, Matt. 2018. *How do we learn to work with intelligent machines? | TED Talk.* https://www.ted.com/talks/matt_beane_how_do_we_learn_to_work_with_intelligent_machines. (Accessed on 03/05/2021).

Bostrom, Nick. 2015. *What happens when our computers get smarter than we are? | TED Talk.* https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are. (Accessed on 03/05/2021).

Citron, Danielle. 2019. *How deepfakes undermine truth and threaten democracy | TED Talk.* https://www.ted.com/talks/danielle_citron_how_deepfakes_undermine_truth_and_threaten_democracy. (Accessed on 03/05/2021).

Hoffman, Guy. 2013. *Robots with soul | TED Talk.* https://www.ted.com/talks/guy_hoffman_robots_with_soul. (Accessed on 03/05/2021).

Kumar, Vijay. 2012. *Robots that fly … and cooperate | TED Talk.* https://www.ted.com/talks/vijay_kumar_robots_that_fly_and_cooperate. (Accessed on 03/05/2021).

Kurzweil, Ray. 2014. *Get ready for hybrid thinking | TED Talk.* https://www.ted.com/talks/ray_kurzweil_get_ready_for_hybrid_thinking. (Accessed on 03/05/2021).

Lee, Kai-Fu. 2018. *How AI can save our humanity | TED Talk.* https://www.ted.com/talks/kai_fu_lee_how_ai_can_save_our_humanity. (Accessed on 03/03/2021).

Meier, Tabea, Ryan L Boyd, Matthias R Mehl, Anne Milek, James W Pennebaker, Mike Martin, Markus Wolf & Andrea B Horn. 2020. Stereotyping in the digital age: male language is "ingenious", female language is "beautiful"–and popular. *PloS one* 15(12). e0243637.

Ordikhani-Seyedlar, Mehdi. 2017. *What happens in your brain when you pay attention? | TED Talk.* https://www.ted.com/talks/mehdi_ordikhani_seyedlar_what_happens_in_your_brain_when_you_pay_attention. (Accessed on 03/05/2021).

Romanelli, Frank, Jeff Cain & Patrick J McNamara. 2014. Should TED Talks be teaching us something? *American Journal of Pharmaceutical Education* 78(6).

Sketch Engine. 2021. *Create and search a text corpus | Sketch Engine.* https://www.sketchengine.eu/. (Accessed on 01/30/2021).

Sugimoto, Cassidy R, Mike Thelwall, Vincent Larivière, Andrew Tsou, Philippe Mongeon & Benoit Macaluso. 2013. Scientists popularizing science: characteristics and impact of TED Talk presenters. *PloS one* 8(4). e62403.

Ted.com. 2021a. *History of TED | Our Organization | About | TED*. https://www.ted.com/about/our-organization/history-of-ted. (Accessed on 01/30/2021).

Ted.com. 2021b. *Our organization | About | TED*. https://www.ted.com/about/our-organization. (Accessed on 01/29/2021).

Ted.com. 2021c. *TED Talks*. https://www.ted.com/talks. (Accessed on 01/30/2021).

Ted.com. 2021d. *TED Talks on AI*. https://www.ted.com/talks?topics%5B%5D=AI. (Accessed on 01/30/2021).

Ted.com. 2021e. *TEDx Rules | Before you start | Organize a local TEDx event | Participate | TED*. https://www.ted.com/participate/organize-a-local-tedx-event/before-you-start/tedx-rules. (Accessed on 03/05/2021).

Tufekci, Zeynep. 2017. *We're building a dystopia just to make people click on ads | TED Talk*. https://www.ted.com/talks/zeynep_tufekci_we_re_building_a_dystopia_just_to_make_people_click_on_ads. (Accessed on 03/05/2021).

Turkle, Sherry. 2012. *Connected, but alone? | TED Talk*. https://www.ted.com/talks/sherry_turkle_connected_but_alone. (Accessed on 03/05/2021).

Wingrove, Peter. 2017. How suitable are TED Talks for academic listening? *Journal of English for Academic Purposes* 30. 79–95.

Chapter 9

How is AI perceived amongst experts of different disciplines?

Maximilian Kalcher

People from the many different academic disciplines and professional fields seem to have a wide range of opinions when discussing AI. Considering those points of view could help one understand how AI will be shaped in the future. The following chapter discusses the views experts of different fields have on AI, through the analysis of podcasts with distinguished guests.

Keywords: Podcasts | Guests | AI | Interview | Opinions

1 Introduction

As of the current year (2021), the discussion around AI in the media has increased and taken many different forms. One of which is podcasting, where one or more recurring hosts normally participate in a discussion about a certain topic. In some podcasts, guests are invited to provide for additional information on a subject or take part in an exchange of opinions. Podcasts have shown to give listeners the ability to dive deep into topics and interesting content as if they were just listening to a long conversation.

This chapter is ordered into five main parts. First of all, the source of information for this chapter, the Lex Fridman Podcast, will be introduced. Then the methodology will be presented before showcasing the guests' opinions in the main body. And, to finalize this chapter, the results will be shown and a conclusion will be drawn.

2 Source

The Lex Fridman Podcast, formally known as the artificial intelligence podcast is a podcast run by the MIT (Massachusetts Institute of Technology) autonomous vehicle and deep learning researcher Lex Fridman. He started the podcast with the intention to talk with renowned MIT professors as part of his taught course 6.S099 ”Artificial General Intelligence”. As more and more people from outside of the institution started watching, and himself not having access to this level of discourse at MIT before, he started broadcasting these podcasts regularly and openly. Nowadays it has become a very well-known podcast which he uploads on his YouTube-Channel, Spotify, amongst other streaming platforms, with a combined over 10,000,000 monthly listener count and guests who are not only professors but also numerous distinguished people in many different fields. Although the name of the podcast has changed to provide for a less constrained discussion of topics, the main idea still involves the discourse around AI.

3 Methodology

For this chapter, each podcast guest was carefully selected based on the following criteria:

1. Having an interesting and unique opinion or approach to AI
2. Representing an academic or professional field that differs from the rest (this also applies to fields with the same parent categories, i.e. software developer and machine learning engineer would both fall under computer science and not be qualified)
3. Adding diversity to the current pool of people (i.e. different background, ethnicity, gender, etc.)

While the majority of guests on the podcasts come from computer-science-related fields with prior and even technical knowledge of AI, the current guests for this chapter were chosen mostly for their indirect relation to it. Therefore, non-recurring guests from fields that differed from computer science were picked. These non-recurring guests were chosen based on the scarcity of the appearance on the podcast, taking for example the chess category, for which there is only one podcast out of the current 168 total aired. The following list shows the fields that were used for this chapter followed by their total number of categorical appearances on the podcast out of the current number of aired podcasts, representing a unique opinion in this context:

Chess (1/168)

Space Travel (6/168)

Entrepreneurship (16/168)

Robotics (7/168)

Economics (3/168)

Comedy (2/168)

Social Media (3/168)

Ethics (4/168)

One of each of those podcasts was transcribed from audio to text via Otter.ai, a free transcription service. This was done to organize each guest according to abstracts of text that stated the content of their discussion around AI, their opinion and also other important key information points that would help for the analysis later on. An important thing to note is that even though the guests presented opinions taking a stance in either direction, a podcast segment is for the most part not sufficient to gather the entirety of the guests' opinion. Therefore, the analyzed segments did not entail other opinions that the individual might have, just the pure substance of their respective podcast occurrence.

4 Podcasts

In this part, each subsection is presented as paraphrases and citations representing the thought of one expert. The title refers to the name of the guest followed by the podcast number and the year of publishing.

4.1 Garry Kasparov (#46, 2019)

Garry Kasparov is considered by many to be the greatest chess player of all time. From 1986 to his retirement in 2005, he dominated the chess world being the world's best player in international ratings for most of those 19 years. While he has many historical matches against human chess players, he might be remembered for one that was not.

In the year 1997, Kasparov played the most important match of his career; against a machine. It was the first time a chess grand-master who was considered the best, was beaten by an AI trained for chess: IBM's Deep Blue. It was not only the first time AI beat a world champion at chess, but it was also the first time Kasparov had lost a professional competitive match in his career, obsessing him to an entire year of ruminating and analyzing this one fatal loss (*Garry Kasparov: Chess, Deep Blue, AI, and Putin – Lex Fridman Podcast #46* 2019).

"I was angry. But look, it was 22 years ago, it's water under the bridge", Kasparov says thinking back to the match today. The big mistake in his opinion was labeling chess as the game of pinnacle human intelligence, because at the end of the day, it is just a game. It's a game and all the machine had to do in this game was just to make fewer mistakes, not solve the game because the game cannot be solved. He explains that the AI was not more intelligent than him during the match, it just found a way by doing something not usual amongst high ratings, capitalized on his mistakes and won. But, after learning the way the machine operated, he was not really impressed. Kasparov says that these early AI programs were made for so-called "closed systems"[1] should not be considered AI, since they would not resemble intelligence, just "brute force"[2]. Only the new programs, namely AlphaZero should belong in that category since they operate with machine-produced knowledge. When looking at the chess games AlphaZero produced, it was intriguing for Kasparov. The program corrected prior game errors and used the knowledge in other future games. But even then, the program had to be tweaked and adjusted by humans.

This is why Kasparov thinks that at the end of the day humans are still flexible enough. He believes we need to recognize our role in the collaboration with machines in the future. Only then will it be beneficial to everyone, when we are not only a worthy opponent for machines but are also ready to lose against them for the sake of progress.

His initial victories and eventual loss to Deep Blue captivated the imagination of the world, leaving people with the question; what role would AI play in the future of civilization? His historic match inspired an entire generation of AI researchers to this day.

4.2 Dava Newman (#51, 2019)

Dava Newman is the Apollo program professor at MIT and the former deputy administrator of NASA (National Aeronautics and Space Administration), and has

[1] A system with a finite set of rules and states. (*Garry Kasparov: Chess, Deep Blue, AI, and Putin – Lex Fridman Podcast #46* 2019))

[2] Methods of solving a problem that relies on sheer computing power and trying every possibility rather than advanced techniques to improve efficiency. (FreeCodeCamp 2020)

been a principal investigator on four space flight missions. Her research interests lie in aerospace biomedical engineering and investigating human performance in varying gravity environments. She has designed, engineered and built advanced spacesuit technology. Although her job does not seem to have a direct connection with AI, it is surprising how important and reliable AI has become for space travel.

When asked about the current role of AI for space exploration, Newman explains the necessity for fully autonomous systems that humans need in order to endure space exploration for the long term and large distances. One big challenge is the design. Fully autonomous technologically equipped people will have to work in real-time without mission control, and all of the systems here on earth being available for assistance, since a 20-minute delay between astronauts and the central station is simply too long. This is why Newman and her team plan on testing equipment out on the moon, also considered the "Proving Ground for testing technologies" (*Dava Newman: Space Exploration, Space Suits, and Life on Mars – Lex Fridman Podcast #51* 2019) for equipment with only a 3-4 second delay once the time is due, before tackling bigger explorations like Mars.

For Newman, the key to exploration are fully robust autonomous systems. She is certain that we, as a civilization, will overcome very difficult challenges in the near future using AI, although humans will always be in the loop some way or another. For the time being, Newman and her research team use AI with terabytes of data given by thousands of satellites to study and predict the earth's temperature and ocean changes. "This next decade, it is urgent that we take care of our own spaceship, that is spaceship earth (before exploring outwards)" (*Dava Newman: Space Exploration, Space Suits, and Life on Mars – Lex Fridman Podcast #51* 2019).

4.3 Elon Musk (#49, 2019)

For many people the entrepreneur of our century, and shortly the newly richest man in the world, Elon Musk, was also brought on the podcasts to discuss AI with Lex. Many people see Elon Musk as the face of technology, leading an electric vehicle company, a space exploration company, implanting chips in brains, building tunnels for fast travel, satellites, solar panels. The fact that Musk publicly discusses concerns about AI is a little surprising, since AI programs or algorithms are used in almost all of his key companies. A special one of which is Tesla's Autopilot, which aims to take on level 5 autonomy [3] by the end of 2022

[3] Level 5 capable vehicles should be able to monitor and maneuver through all road conditions and require no human interventions whatsoever, eliminating the need for a steering wheel and

by using its signature recurrent deep neural networks (Döpfner 2020).

"It is important to have a referee that is serving the public interest" (*Elon Musk: Neuralink, AI, Autopilot, and the Pale Blue Dot – Lex Fridman Podcast #49* 2019), he says during his second time on the podcast. The concerns expressed on the podcast involve AI safety and regulation. He believes there needs to be a (US) government agency overseeing anything related to AI to confirm that it does not represent a public safety risk, just like there are regulatory agencies for food and drugs, automotive safety, etc. Even if it is a little far fetched, Musk does not rule out the possibility of the "singularity", also known as the point of no return for superintelligent AI systems, and wishes for people to steer the notion for a more positive outcome than a negative one. For more in-depth information on Elon Musk's opinions on AI outside of his podcast occurrence, see chapter 10.

4.4 Kate Darling (#98, 2020)

People say that the one thing you cannot teach a computer is emotion, Kate Darling is one of the key people trying very hard to solve this. As a researcher at MIT interested in social robotics and ethics, she explores the emotional connection between human beings and life-like machines. She is also a caretaker of several domestic robots.

When Lex asked if it was possible to fall in love with robots, she immediately responded: "Yes, definitely". But then followed up by explaining that she would rather not compare robot AI companions to humans but rather to pets. Although these AI companions should help alleviate loneliness, they should not necessarily have the same role as humans do, but rather be supplemental in a different way. "While people are constantly worried about robots replacing humans [....], we rarely talk about robots actually filling a hole where there is nothing and what benefit that can provide to people", she says during the interview.

She acknowledges that, even though it is not the main use for them, people will still try to use AI companions to form romantic or even sexual relationships. At the end of the day, she does not rule out the importance of very natural human-like qualities which are necessary for human interaction, like human-like voices, emotions, touch, smell or even just having a body that resembles ours.

What does it take to create a system that resembles humans? How hard is it to create conversational agents? How hard is it to pass the Turing Test[4]? It's all

pedals. (TrueCar, Inc. 2018)

[4] A hypothetical test to clarify the question as to whether computers can think. (Oxford Reference 2021)

about expectation management, she says. She takes for example Sophia, the social humanoid robot developed by Hong Kong-based company Hanson Robotics. Even though the robot is very advanced in a way that it has human-like qualities, a face, is able to hold long conversations and even imitates human gestures and facial expressions, it is still not enough for her to be impressed. Building robots for loneliness is a very difficult challenge, and Darling thinks it will take far more time than people realize for them to be placed in our society. "I think people are a little bit influenced by science fiction and pop-culture, to think that we should be further along than we are" (*Kate Darling: Social Robotics – Lex Fridman Podcast #98* 2020). But all in all, Kate does not see a future without social robot systems in our proximity, however far this future might be.

4.5 Erik Brynjolfsson (#141, 2020)

In podcast #141, Lex talks to Erik Brynjolfsson. He is an economics professor at Stanford and the director of Stanford's Digital Economy Lab, after having previously worked a long time at MIT where he taught and broke down the economics of information. The impact of artificial intelligence and automation on our economy and world as a whole is something worth thinking deeply about. Like with many topics linked with predicting the future of society, it's easy to fall into one of two categories, the utopia or the dystopia.

Brynjolfsson calls himself a "technology optimist" and not really at the "Singularity is near" end of the spectrum, at least in the coming decades. He thinks there are likely to be some significantly improved living standards and some really important progress. The main part to notice is that it takes 10, 20 or 30 years for the existing technology to have profound effects. "Even if nothing new got invented, we would have a few decades of progress", he says and is very excited about that. One of the applications he is most excited about is health care, which is going to lead us to live healthier and therefore wealthier lives.

A very big center of discussion remains the fact that AI is supposed to replace massive amounts of jobs in the future. Brynjolfsson does not think we are going to have the end of work anytime soon. There are just too many things that machines still cannot do, whether it's child care or health care, interacting with people, scientific or artistic work that requires creativity. These are things that, for now, machines are not able to do nearly as well as humans, even just something as mundane as folding laundry. "[…] Many of these I think are going to be years or decades before machines catch up" (*Erik Brynjolfsson: Economics of AI, Social Networks, and Technology – Lex Fridman Podcast #141* 2020), he thinks.

As far as the next big thing for humans according to Brynjolfsson, is working together with AI in a symbiotic kind-of relationship in the workspace. In the next 10, 20, 30 years there will be a big restructuring of society, some people will get wealthier, some will have to learn new skills. Now, says Brynjolfsson, if we were to go back even further into the future, like 50 or 100 years, all bets are off and it's possible for machines to be able to do almost anything a human can. And at this point, we would start to get into an economy of abundance [5], a world where there is really little for humans to do economically better than machines, other than to be human.

4.6 Whitney Cummings (#55, 2019)

A problem that is not thought about thoroughly in terms of robots and AI is gender. What gender should robots of certain use have? Should we have genderless robots? Whitney Cummings is a stand-up comedian, actress, producer, writer and host of her own podcast called "Good for you". In 2019 she also had a Netflix special called "Can I touch it?", which features a robot playing a replica of her. In their conversation, Cummings and Lex touch on the social aspects of robotics and AI.

To Cummings, the idea of genderless robots makes a lot of sense for things like babysitting, given her husband is in the house. Genderless robots should only then be as a helping hand and not as a substitution to a parent, most importantly considering younger children's gender identity development early on. "You know, there are places that I think that genderless makes a lot of sense", she says. "But obviously not in the sex area" (*Whitney Cummings: Comedy, Robotics, Neurology, and Love – Lex Fridman Podcast #55* 2019). Cummings had previously visited a sex-robot factory for a part of her comedy sketch. She recalls that some workers told her that even non-straight people had even found it helpful experimenting with sex robots of the same gender, just to test it out, before moving to humans.

For Cummings, robots are in general the only solution for cleaning after the mess humans leave behind, pointing towards the pollution in the oceans. Also, for general safety hazards: "You know, firefighters are heroes but they're limited in how many times they can run into a fire [...]" (*Whitney Cummings: Comedy, Robotics, Neurology, and Love – Lex Fridman Podcast #55* 2019). They also might be not just helpful for nature-induced hazards, but also keeping humans safe from

[5] A theoretical economic situation in which most goods can be produced in great abundance with minimal human labor needed, so that they become available to all very cheaply or even freely. (Wikipedia contributors 2021)

other humans, especially those more at risk: "[...] we could see them maybe as like, free assistance, help and protection. And then there's sort of another element for me personally, which is maybe more of a female problem."

4.7 Jack Dorsey (#91, 2020)

Almost everyone uses social media, notably one of the platforms most used today due to the easy engagement for people is Twitter. When founding Twitter, Jack Dorsey had always used state of the art data analytic tools which nowadays would fall under the name of machine learning to handle massive amounts of data. Jack has always been a fan of new technology, publicly advocating for decentralization and Bitcoin, usage of big data, and other topics.

Although being a clear optimist about AI, Dorsey had only a short segment on the podcast talking about it. When he was asked what it took to pass the Turing Test in the space of language, he told Lex that where we are now and at least for years out, the combination of machine learning and AI models paired with human discussion depth, nuance and meaning is something very interesting for him. Dorsey finds it important for these intelligent machines to be able to use natural language for self-descriptiveness or for the usage of finding meaning in data sets. He follows this thought with the big problem and risk he has with AI going forward, namely that the field is building more and more black boxes[6] that may lead to a correct but inexplicable result when being used. "[...] And we are trusting them more and more from lending decisions to contact recommendation to driving to health [...]" (*Jack Dorsey: Square, Cryptocurrency, and Artificial Intelligence – Lex Fridman Podcast #91* 2020).

Dorsey also discusses the importance of detection technology being 10 steps ahead of creating technologies, basically a race. A lot of work he deals with currently in his co-founded financial service company Square Inc., is built around identity. Payments ultimately come down to that. And he fears, that using the new DeepFake technology as an example, not only will constructing false identities to accept payments or use faulty credit cards be easier in the future, but also could really damage the security of the state, taking into consideration falsifying passports, identifications, drivers' licenses. Jack believes that specifically with Deepfakes, the detection technology is already lagging behind at this point.

[6] A device, system or object which can be viewed in terms of its inputs and outputs (or transfer characteristics), without any knowledge of its internal workings. (Kenton 2020)

4.8 Ayanna Howard (#66, 2020)

Ayanna Howard is a roboticist, professor at the Georgia-Tech-Institute and director of the human automation systems lab. Her research interest lies in human-robot-interaction, system robots for at-home usage, therapy gaming apps and remote robotic exploration of extreme environments. But just as much as she cares about robots, she also cares about humans.

In their conversation, having worked on semi-autonomous vehicle technology, namely for Tesla, Lex started off by asking about the ethical responsibilities of developers in the world of AI. Every semi-autonomous vehicle navigates using a certain risk function. Meaning there is an objective function in every self-driving car that computes risk probabilities, for example the probability of killing another human being. This function, first of all, has to be low enough to be acceptable on an ethical level to not be a public safety concern, but also has to be high enough for people on the street to respect you and not try to cross the path of the car. Therefore, ultimately a developer has some indirect responsibility in the death of a human being. Howard explains how having a clear ethics system in the AI community is very important: "[...] You can basically say, I'm not going to work on weaponized AI [...], but yet you are programming algorithms that might be used in healthcare that might decide whether this person should get this medication or not. And they don't, and they die".

People can be unconsciously making decisions and are unaware that they have that power when they are coding. Because of this problem, she proposes to really think about responsibilities more than we currently are. Developers should go back to the "early days of developing" (*Ayanna Howard: Human-Robot Interaction & Ethics of Safety-Critical Systems – Lex Fridman Podcast #66* 2020), meaning to not just compile the code and rely on so-called ethical testers, just because they assume their work goes through another process, but to actually be the ethical tester yourself.

Howard hopes for developers to be able to have great responsibility for the amount of power they have. She compares developers with direct impact on society, like the ones developing for semi-autonomous vehicles to surgical doctors, who are mostly given tools to approach every decision with absolute caution and the probability of the patient not surviving, while not going crazy in the process. She believes some of those tools should become available in some form for developers as well, at least in the future, where lives become more directly dependent on AI.

5 Results

So far, all guests have shown to have an interesting approach to AI when it comes to their current work, hopes for the future and general thoughts. In the following, the segments in which the guests talk about AI are classified in three groups, the optimistic, the pessimistic and the neutral group. The optimistic group refers to the guests who talk about how they make use of AI in their work or believe AI will be used more positively than negatively in the future. The pessimistic group is the opposite, either they do not make use of AI or believe that AI will mainly be used for the negative in the future. This group also includes people who are cautious about AI, believing there can be substantial risks to it. And then at last, there is the neutral group who have either have both positive and negative thoughts about AI or don't communicate a clear stance.

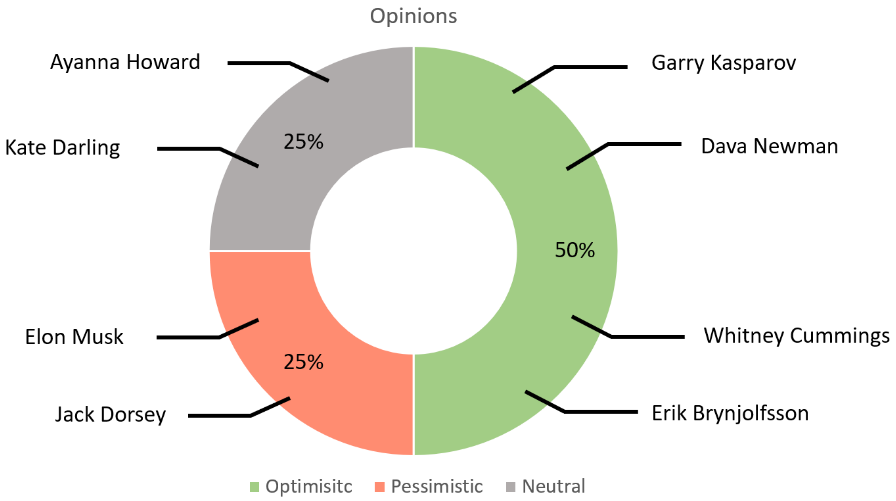


Figure 1: Categorization of the guests' opinions

When taking a look at the chart in Figure 1, four of the eight total chosen guests seem to have an optimistic stance about AI. Two of those eight point out more concerns than hopes and the other two have both an equal positive and negative opinion. As we can gather, the majority of guests have a clear positive view on AI.

Another interesting result of the analysis was the usage of certain keywords in the podcasts given by the keyword finder of the transcription service Otter.ai. Figure 2 shows a bar chart with the most frequently used keywords in the optimistic,

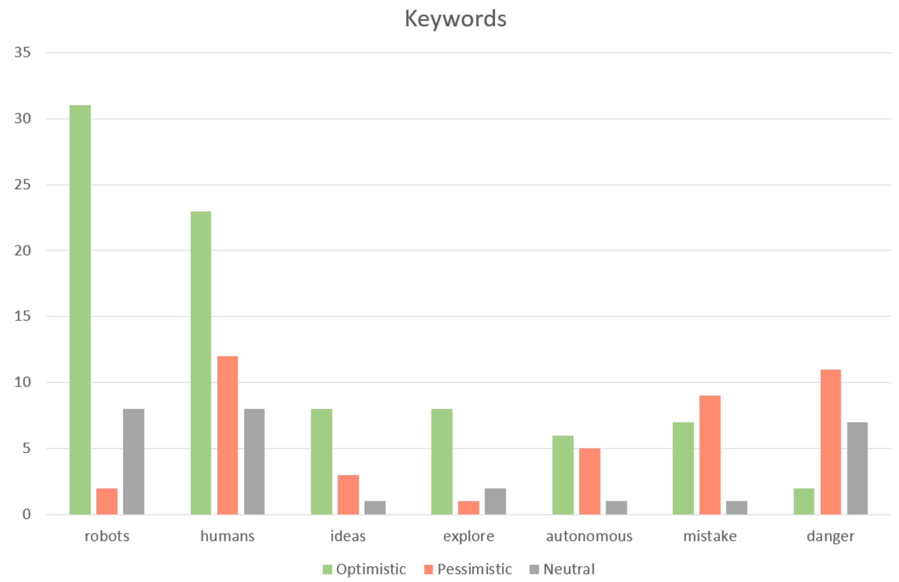


Figure 2: Respective keyword count by groups (Credit: Otter.ai)

the pessimistic and the neutral group accordingly. Noteworthy is for the most part the increasing usage of the words ”robots” and ”humans” in the optimistic group, pointing towards discussions of optimistic unity. These are followed by ”ideas”, ”explore” and ”autonomous”. Moreover, the pessimistic group makes use of words like ”danger”, ”mistake”, and also ”humans”. This can be interpreted as keywords for a more cautionary discussion. The neutral group makes use of a mixture of word counts between the optimistic and pessimistic groups.

6 Conclusion

In conclusion, most guests have shown to make good use of AI in their current work or believe that it will benefit society and be used positively in the future. These guests mostly associate the positive discussion around AI addressing both humans and robots, almost just as much. A big part of this is also the mention of exploration and ideas, including automation.

To finalize the conclusion, public discourse around AI in the form of podcasts has proven to be an effective method of clear information transmission to listeners all around the world. And hopefully, the reader has learned from some distinguished guests that a positive image about AI is present in fields we might not expect right away.

References

Ayanna Howard: Human-Robot Interaction & Ethics of Safety-Critical Systems – Lex Fridman Podcast #66. 2020. Podcast. (Accessed on 10/03/2021). https://youtu.be/J21-7AsUcgM?t=1203.

Dava Newman: Space Exploration, Space Suits, and Life on Mars – Lex Fridman Podcast #51. 2019. Podcast. (Accessed on 10/03/2021). https://youtu.be/2fI6bYnRgSc?t=1921.

Döpfner, Mathias. 2020. *Elon musk reveals Tesla's plan to be at the forefront of a self-driving-car revolution — and why he wants to be buried on mars.* Interview with Elon Musk. https://www.businessinsider.com/elon-musk-interview-axel-springer-tesla-accelerate-advent-of-sustainable-energy?r=DE&IR=T.

Elon Musk: Neuralink, AI, Autopilot, and the Pale Blue Dot – Lex Fridman Podcast #49. 2019. Podcast. (Accessed on 10/03/2021). https://youtu.be/smK9dgdTl40.

Erik Brynjolfsson: Economics of AI, Social Networks, and Technology – Lex Fridman Podcast #141. 2020. Podcast. (Accessed on 10/03/2021). https://youtu.be/NOReE-3EBhI?t=3191.

FreeCodeCamp. 2020. Definition of brute force computing. https://cutt.ly/0xryWZP.

Garry Kasparov: Chess, Deep Blue, AI, and Putin – Lex Fridman Podcast #46. 2019. Podcast. (Accessed on 10/03/2021). https://youtu.be/8RVa0THWUWw?t=1386.

Jack Dorsey: Square, Cryptocurrency, and Artificial Intelligence – Lex Fridman Podcast #91. 2020. Podcast. (Accessed on 10/03/2021). https://youtu.be/60KJz1BVTyU?t=1531.

Kate Darling: Social Robotics – Lex Fridman Podcast #98. 2020. Podcast. (Accessed on 10/03/2021). https://youtu.be/7KTbEn7PiaY?t=1227.

Kenton, Will. 2020. Definition of Black Box. https://www.investopedia.com/terms/b/blackbox.asp.

Oxford Reference. 2021. Definition of Turing Test. https://www.oxfordreference.com/view/10.1093/oi/authority.20110803110135741.

TrueCar, Inc. 2018. The 5 levels of car autonomy. https://www.truecar.com/blog/5-levels-autonomous-vehicles/.

Whitney Cummings: Comedy, Robotics, Neurology, and Love – Lex Fridman Podcast #55. 2019. Podcast. (Accessed on 10/03/2021). https://youtu.be/0-3kw5BEKB8?t=282.

Wikipedia contributors. 2021. Definition of post scarcity economy. https://en.wikipedia.org/wiki/Post-scarcity_economy.

Chapter 10

How does Elon Musk portray the dangers and the future of AI?

Alina Deuschle & Joline Janz

In this chapter, we will focus on how Elon Musk – who is a highly influential tech entrepreneur – portrays the future and danger of Artificial Intelligence, also referred to as 'AI'. As someone very prominent and highly influential within the realm of Artificial Intelligence, constantly portraying his future visions and possible dangers he appeared to be a suitable source for discussing AI in public discourse. Therefore, this chapter will not portray different opinions on the matter at hand, but rather concentrates on Elon Musk being a very outspoken entrepreneur who impacts the public discourse in various ways through his visions and his inventions. The aim of this chapter is to clearly present the thoughts and concerns about AI of Elon Musk. To gain a better understanding of why exactly he plays such an important role within this field of expertise, we will focus on him as a successful entrepreneur first. Afterward, the chapter gives insight into the methodology used during the examination process. Moving on, two scenarios that Musk envisions when it comes to the future of AI are outlined – one of which is a benign scenario, whereas the other is considered malignant. In the conclusion the results will be discussed and explained why they underline Elon Musk's position as a representative of the public discourse of Artificial Intelligence.

Keywords: Elon Musk | Artificial Intelligence | Future | Danger

1 Introduction

In this paper, Elon Musk's role in the public discourse of Artificial Intelligence is discussed. One might wonder why the focus lies on a single person when the aim is to discuss the public discourse on the issue. One naturally would expect a discussion of the understanding of Artificial Intelligence on a broader scale. So how does it come that – according to the headline of this paper – it seems to be concerned with only one public figure, namely Elon Musk? The introduction aims to provide an answer to this question.

Before diving into the discussion, it is of vital importance to portray Musk's character as well as outlining his career path, since those aspects have immensely shaped his opinion on Artificial Intelligence.

2 Who is Elon Musk?

Born in South Africa in 1971, he showed an interest in technology already at an early age. At the age of twelve, Musk sold his first video game to a computer magazine. Due to his unwillingness to support apartheid through compulsory military service, Musk decided to move to his mother's home country, namely Canada, where he studied at Queen's University in Kingston, Ontario. Musk always planned to move to the United States in the hope of greater economic opportunities. Therefore, it was a natural decision for him to transfer to the University of Pennsylvania, Philadelphia, to further pursue his Bachelor's degree in physics and economics. After his studies, Musk realised the great potential that the World Wide Web bears and decided to quit his planned further studies in physics at Stanford University in California. The foundation for Musk's career was already laid while he was still studying at the University of Pennsylvania. The Zip2 company, which was sold in 1999 to a computer manufacturer for 307 million dollars. Following this success, he founded an online financial service that became well known as the online money transferring platform PayPal.

He is now CEO and investor of multiple companies, of which – amongst others – the most prominent are SpaceX, Tesla, and Neuralink. Through his successful entrepreneurship, he has become one of the richest men in the world. Not only that, but also his outspokenness about the visions of his companies and also possible scenarios of the future of Artificial Intelligence render him a very influential person.

Therefore, Musk is a very acclaimed guest of interviewers and researchers from a versatile field of interests. Additionally, he has spoken to several instances of the US government about his concerns when it comes to the future of AI. It appears as if Musk considers himself somewhat responsible to educate the public about the dangers AI could bring with. He has reached out to Barack Obama to share his thoughts on the issue at stake. Therefore, it becomes apparent why Elon Musk is of high influence when it comes to the public discourse of AI.(Gregersen 2020)

3 Methodology

To investigate Musk's views on AI, we decided to focus on using videos uploaded on the video hosting platform YouTube. Our main reason for choosing YouTube is the feature that everyone is able to upload videos on the platform. This will allow us to access interviews from different kinds of hosts and therefore would contribute to get a better grasp of the public discourse of Artificial Intelligence.

To better match the visitor's preference, YouTube offers the possibility to adjust various filters and thereby restrict the search query. This tool enabled us to specify the resulting video suggestions according to our requests as well as it helped us to create a manageable amount of sources to evaluate.

After searching for the keywords "Elon Musk Artificial Intelligence", additionally we narrowed our sources to include only those videos longer than 20 minutes and those watched by at least one million users. Our reason for setting the first filter were that we decided to focus on full interviews since short extracts do not include the full picture of a conversation and can lead to false conclusions. Regarding the second filter, we only wanted to include those videos that have sparked a considerable amount of attention and are therefore influential on the public discourse. Other than that, we included those videos directly featuring Elon Musk, since we wanted to base our analysis on direct quotes. Thereby, we excluded any kind of reaction videos.

Seventeen videos matched our filter setting, of which we had to exclude five videos due to not containing direct speech from Elon Musk. Since we could not retrieve the scripts of all the videos, we created notes ourselves. After we finished watching all the videos, we restructured our notes in a way that contrasts the arguments in favour of AI with the arguments that are more negative with regards to Artificial Intelligence.

4 Discussion

4.1 Overview of the Sources

In the first part of the discussion, the eleven videos that met the requirements will be portrayed in order of the number of views. Through this a summary of the main topics Musk focuses on in his interviews is established. Moving on, we will focus on his position and future visions regarding the topic of Artificial Intelligence.

The two videos with the most clicks as of March 2021, namely more than 41 and 21 million, are both interviews held by Joe Rogan within his podcast "The Joe

Rogan Experience". The latter was held in May 2020, whereas the most popular one was recorded in September 2018. Joe Rogan – who is an American comedian and television host – created his podcast in 2009. His podcast is one of the world's most well-known, featuring a wide area of guests.

Following the two full podcast episodes, the next two most-watched videos are cut-outs of Joe's and Elon's conversation. The first, which is an excerpt of the first podcast episode the two recorded was clicked over 7.5 million, the second, which portrays a conversation originally taking place in the second podcast that was recorded has 3.6 million clicks.

Following with more than 2.8 million clicks is a debate held between Jack Ma and Elon Musk in Shanghai at the "World Artificial Intelligence Conference" in the summer of 2019. Jack Ma, who is a Alibaba Group co-founder and executive chairman, and Elon Musk discussed several questions regarding the topic of Artificial Intelligence and shared different thoughts and visions.

Another video with more than 2.3 million clicks shows Elon Musk being interviewed at the Code Conference in 2016. The Code Conference is held by the journalists Walt Mossberg and Kara Swisher and features several interviews with experts in the field of digital technology on the issue of current and future impacts of the matter at stake.

Next up is an excerpt of an interview between Elon Musk and Jonathan Nolan. The conversation was held in October 2018 and has been clicked over 1.8 million times.

The next video that matched the filter requirements has been clicked on more than 1.5 million times and shows one of the two interviews between Elon Musk and Lex Friedman. Next to being a podcast host, Lex Friedman works at MIT in different areas of research including Artificial Intelligence, Autonomous Vehicles, and Machine Learning. He decided to start his podcast, which uploads new episodes in a biweekly manner, to share findings in the topics concerning his areas of research. Within this realm, he has also interviewed Elon Musk twice, first in April 2019 and a second time just half a year later in November 2019. Both videos have roughly the same amount of views, however the first one did get a little bit more attention.

Another video that also reached a similar amount of clicks is the Neuralink Launch Event in July 2019. Neuralink is one of the companies Musk is involved in, it is interested in finding a way to merge biological intelligence with machine intelligence. Even though that is the company's ultimate goal, Musk also envisions ways to relieve several brain diseases. This chapter will focus more on the concept and hopes of Neuralink later on.

Lastly, a video with more than 1.2 million clicks shows an interview of Elon Musk at the World Government Summit convention in February 2017. The interview was held by the Minister of Cabinet Affairs of the United Arab Emirates, Mohammad Abdullah Al Gergawi.

All of the videos featured above discuss a similar range of topics, which will be summarized in the following section. The main topics discussed – in almost every video listed above – contain Artificial Intelligence, a sustainable way of living for the human race – including the possibility to become a multi-planetary species, as well as insights into Musk's current projects and visions. Irrespective of the broad range of topics at hand, the focus of this chapter will mainly lie on his opinions and visions concerning AI, its impact on the future, as well as possible projects that are directly connected to the matter at hand.

4.2 The Future of Artificial Intelligence

Throughout the videos Elon Musk portrays two possible future scenarios concerning the development of Artificial Intelligence – the first one can be considered as a benign scenario, whereas – on the contrary – the second one paints a much more pessimistic view and will therefore be referred to as the malignant scenario.

Elon Musk emphasises the immense rate of advancement throughout almost all of the interviews. He urges the importance of the matter by stating: "The rate of improvement is really dramatic [...]" (**03:31**, Nolan 2018). Due to this increase in advancement, Musk predicts that humanity will reach the point of singularity sooner or later. Singularity is a term mostly used with respect to black holes. The centre of a black hole is referred to as singularity, a point where all matter is compressed to an infinitely tiny point. Singularity is more of an abstract understanding, it is not exactly clear what happens once the point is reached (NASA 2018). He refers to the rapid improvement of Artificial Intelligence as a countdown that is out of control. Once a certain point is reached, namely the point of singularity, it's development is hard to predict, similar to a black hole (Rogan 2018).

By that, Elon Musk means that once Artificial Intelligence has reached a certain point of advancement and therefore reaches singularity it will be out of human control. "It could be terrible and it could be great. It is not clear, but one thing is for sure. We will not control it."(**24:00**, Rogan 2018) Elon Musk therefore infers that a possible future scenario of AI does not necessarily have to be bad. However, he states: "I am concerned about certain directions that AI could take that would be not good for the future […], not all AI futures are benign" (**41:44**, Swisher & Mossberg 2016).

Elon Musk is concerned with people's tendency to underestimate the power and amount of intelligence AI systems will reach in the future. The aforementioned immense rate of improvement will lead to a shift in the amount of intelligence machines entail compared to humans. He refers to the human species as a "biological boot loader for AI" (**17:25**, Rogan 2018), meaning we are constantly feeding AI and therefore help it to become more intelligent. Musk states: "We are building progressively greater intelligence. And the percentage of intelligence that is not human is increasing. And eventually we will represent a small percentage of intelligence." (**17:36**, Rogan 2018)

He also thinks that one of the biggest mistakes is to assume that machines will not become smarter than humans. According to Musk, it is not up for debate that machines will outsmart the human race, which will not be able to compete in the future (*Elon Musk and Jack Ma hold debate in Shanghai* 2018). He emphasises the rate of improvement and argues that it will outpace the human ability to understand the AI (*Elon Musk and Jack Ma hold debate in Shanghai* 2018).

Another point Musk brings up is the growing temptation to use AI as a weapon. According to Musk, it is not only the danger AI brings with it by itself, but the fact that humans can use it against each other (Rogan 2018). He states: "I think the danger of AI is much greater than the danger of nuclear warheads – by a lot. […] Mark my words, AI is far more dangerous than nukes, far, so why do we have no regulatory oversight" (**04:30**, Nolan 2018).

Musk hereby introduces another important point, namely the necessity to regulate the usage of AI. Musk argues that regulations generally work very slowly (**24:00**, Rogan 2018) and he therefore emphasises that it is important to start the process sooner rather than later and to slow down the improvement of AI, otherwise it might be too late (Rogan 2018).

Even though Elon Musk sees many challenges the future of AI bears, he also envisions a benign scenario – a scenario where human and machine live in symbiosis. One thing he repeatedly states is: "If you can't beat it, join it." (**24:35**, Rogan 2018). According to Musk, it is for sure that there will be a point where AI will outsmart the human race. He therefore seeks a way that still allows the human race to compete.

Musk argues that humans have already extended their intelligence through technology. One's phone – as well as computer – can be seen as an external source of intelligence. Musk wants to extend this concept by adding some sort of intelligent layer to the brain. "I think one of the solutions, the solution that seems maybe the best one, is to have an AI layer, if you think about it, you got your limbic system, your cortex and then a digital layer, sort of a third layer above the cortex." (**58:00**, Swisher & Mossberg 2016) Musk argues that the way in which the limbic system and the cortex communicate could be mimicked and a potential third digital layer could work symbiotically with the rest (Swisher & Mossberg 2016).

Musk therefore co-founded the company Neuralink in July 2016. The company's goal is to find a way to create human-machine communication directly in the brain.

To achieve this, a robotic system will be used that can reliably insert Micron-scale threads into areas of the brain responsible for controlling movement. These threads contain electrodes that are in a further step connected to the neural implant, called the "Link". Due to the thread's fineness and flexibility the surgery cannot be performed by a human hand. The overall goal is to support patients with paralysis to regain their independence and control through devices, such as our phones and computers. The Link will open up the possibility to operate devices from everywhere in the world. This invention is the origin of an innovative kind of human brain interface. (Neuralink 2021) The company's future vision concerning the Link regards the augmentation of communication channels with the human brain, by accessing more interconnected brain areas. Furthermore, it is expected that by further refining the technology this tool will enable us to treat diverse neurological disorders, recover movement and sensory function, and extend our way to interact. (Neuralink 2019)

5 Conclusion

Elon Musk is overall very outspoken about the possible dangers of Artificial Intelligence in the future. However, he tries to retain an optimistic stance towards the topic by bringing the vision to life that could enable the human race to compete with the super-intelligence of AI. His opinions and thoughts have a high influence on the overall stance on the issue at hand since his successful entrepreneurship allows him to share his thoughts on a large scale.

6 Appendix – Filtered Youtube Videos

Title	**Interviewer**	**Year**	**Views**
Joe Rogan Experience #1169 – Elon Musk	Joe Rogan	07.09.2018	41,813,462
Joe Rogan Experience #1470 – Elon Musk	Joe Rogan	07.05.2020	21,137,504
Joe Rogan – Elon Musk on Artificial Intelligence	Joe Rogan	07.09.2018	5,750,803
Elon Musk Reveals New Details About Neuralink, His Brain Implant Technology	Joe Rogan	07.05.2020	3,609,434
Artificial Intelligence: Mankind's Last Invention	–	05.10.2018	3,356,905
When Elon Musk Realized China's Richest Man Is A Dope (Jack Ma)	–	26.09.2019	3,228,424
Jack Ma and Elon Musk hold debate in Shanghai	–	29.08.2019	2,882,683
Why AI will probably kill us all.	–	05.03.2017	2,365,537
Elon Musk | Full interview | Code Conference 2016	Kara Swisher & Walt Mossberg	02.06.2016	2,361,933
Elon Musk's Last Warning About Artificial Intelligence	Jonathan Nolan	03.10.2018	1,829,404
Elon Musk: Tesla Autopilot | Lex Fridman Podcast #18	Lex Fridman	12.04.2019	1,585,696
Neuralink Launch Event	–	17.07.2019	1,573,923
Elon Musk: Neuralink, AI, Autopilot, and the Pale Blue Dot | Lex Fridman Podcast #49	Lex Fridman	12.11.2011	1,559,209
Joe Rogan Talks Artificial Intelligence with a Yale Professor	Joe Rogan	28.03.2019	1,539,635
Billionaires Jack Ma vs. Elon Musk debate in Shanghai at World Artificial Intelligence	–	31.08.2019	1,526,900
WGS17 Session: A Conversation with Elon Musk	H.I. Muhammad Al Gergawi	15.02.2017	1,220,005
ELON MUSK – Tesla | SpaceX | Solar City | Open AI | Boring Company | Paypal | The Story So Far	–	18.07.2019	1,211,641

All Videos accessed on March, 6th 2020.

References

Elon Musk and Jack Ma hold debate in Shanghai. 2018. https://www.youtube.com/watch?v=f3lUEnMaiAU&t=1s. (Accessed on 06/03/2021).

Gregersen, Erik. 2020. *Elon Musk.* https://www.britannica.com/biography/Elon-Musk. (Accessed on: 07/03/2021).

NASA. 2018. *What is a black hole?* https://www.nasa.gov/audience/forstudents/k-4/stories/nasa-knows/what-is-a-black-hole-k4.html. (Accessed on 08/03/2021).

Neuralink. 2019. *Neuralink launch event.* https://www.youtube.com/watch?v=r-vbh3t7WVI&t=5113s. (Accessed on 06/03/2021).

Neuralink. 2021. *Breakthrough technology for the brain.* https://neuralink.com. (Accessed on: 07/03/2021).

Nolan, Jonathan. 2018. *Elon Musk's last warning about artificial intelligence.* https://www.youtube.com/watch?v=B-Osn1gMNtw&t=43s. (Accessed on 06/03/2021).

Rogan, Joe. 2018. *Joe Rogan Experience #1169 - Elon Musk.* https://www.youtube.com/watch?v=ycPr5-27vSI&t=1069s. (Accessed on 03/03/2021).

Swisher, Kara & Walt Mossberg. 2016. *Elon Musk | full interview | Code Conference 2016.* https://www.youtube.com/watch?v=wsixsRI-Sz4&t=2521s. (Accessed on 06/03/2021).

Chapter 11

How is AI in healthcare perceived by physicians in 2020? – A qualitative analysis of recent opinions in YouTube videos

Janine Reichmann & Ali Jandaghi

With artificial intelligence becoming a major player in different economic and technical sectors, it also brings change to medicine and the healthcare system. Many AI-based techniques are already in use which leads to a rising debate of pros and cons of AI in medicine conducted by many people of different professions. However, as especially physicians, e.g. in clinics, have to work with AI, their opinions are quite important in the ongoing discussion.

Therefore, the purpose of this research is to analyze how physicians actually perceive AI in medicine nowadays, compare their opinions and get an insight into their view. To achieve that, various recent YouTube videos of physicians talking about this topic were analyzed with a qualitative approach using certain criteria to divide the discussion into several subtopics. The results showed that the majority of physicians have a positive attitude toward AI and find more strengths than weaknesses. The fear of AI stealing their job did not seem to be from great importance. However, the aforementioned weaknesses must be taken into account.

Keywords: Artificial Intelligence | Physician | Healthcare | Medicine | Perception

1 Introduction

In the course of increasing digitalization, artificial intelligence plays a key role in big data processing, data analysis, automatic reasoning and many more. A rising topic of the last years was the application of AI in medicine and healthcare systems, which led to a growing debate about the corresponding challenges and benefits which come with the development. Additionally, the interest in AI and the willingness of various organizations and countries to actually use and implement these technologies are constantly increasing (Khoroshevsky et al. 2020).

Instead of only one unique definition for artificial intelligence, there are various to be found. However, in general AI is able to learn and solve problems like a human would solve it by applying different methods e.g. Machine Learning (ML), Deep Learning or Neural Networks. When used in medicine or healthcare, the problems AI has to solve are clinical cases about diagnosing or treating patients. Buch et al. (2018: 143) describe how AI can be implemented in the medical sector:

> For example, neural networks represent data through vast numbers of interconnected neurons in a similar fashion to the human brain. This allows ML systems to approach complex problem solving just as a clinician might — by carefully weighing evidence to reach reasoned conclusions. However, unlike a single clinician, these systems can simultaneously observe and rapidly process an almost limitless number of inputs.

Presently, AI is frequently used in various fields of medicine. For example, it can serve as a helpful diagnostic tool in various cancer treatments: An AI algorithm can increase the probability to detect breast cancer (Watanabe et al. 2019) or lung and liver cancer at a much earlier stage (Patel et al. 2020). Normally, lung and liver cancer are detected at a very advanced stadium, that is why AI can reduce the death rate rapidly (Patel et al. 2020).

Additionally, artificial intelligence offers many possibilities in the field of cardiology and especially cardiac imaging. This field was one of the early adopters of AI, so that several techniques are already in use (Dilsizian & Siegel 2014). On the one hand, as in cancer detection, AI serves as a diagnostic tool in cardiac imaging allowing physicians to interpret more images correctly (DePuey et al. 1989) by comparing an image to a large normal data base or to get support for the diagnosis and treatment of coronary artery disease and the detection of arrhythmia (Dilsizian & Siegel 2014). On the other hand, as stated by Dilsizian & Siegel (2014: 5), "AI serves as an image enhancement technique rather than as a diagnostic tool" when used in a Positron Emission Tomography (PET) scan to highlight abnormalities. So, depending on the use case, AI can serve either as a diagnostic tool or as an image enhancement technique.

Artificial intelligence can also be used to discover, develop and optimize drugs e.g. for cancer patients. Up until now, the algorithms can successfully predict drug behavior, design drug combinations, modulate "multidrug dosing using only a patient's data" (Ho 2020b: 983) and design a new drug compound in only 21 days while a conventional approach would need a whole year for this (Ho 2020b). Overall, AI is able to reduce healthcare costs, support for physicians in different fields of medicine and it provides more access to personalized medicine.

However, there is also criticism based on ethical questions and challenges that emerge when utilizing AI as support in the healthcare sector. For example, the question how AI in medicine shapes human behavior is widely discussed. That means that a physician's decision making and diagnostic skills may be changed or weakened because they rely too much on AI (Currie et al. 2020). Moreover, this problem leads to the question of liability when a false diagnosis was made or a wrong medication was given to a patient. Currie et al. (2020: 749) suggest the following to avoid this challenge: "Clear guidelines need to be developed when decisions are made based on the output of AI in terms of ethical and legal responsibility […]". Furthermore, the discourse about AI often brings about the criticism of its black box property, which refers to the lack of transparency that makes it impossible for humans to understand the decision-making process behind the outcome (Bathaee 2017).

Undeniably, there is a vast discussion about AI's abilities and what it should or should not do. However, in literature and in the media, the opinions of physicians are often left out completely although they are the ones that will actually work with new AI technologies once they are implemented. Therefore, this leads to the research question *"How is AI in healthcare perceived by physicians in 2020?"* where we collect and compare different opinions of physicians to see how they perceive the rising trend of utilizing AI in medicine and healthcare.

2 Methods

To answer the research question and to get an insight into the physicians' opinions, YouTube videos by or with physicians were chosen. In the following, they were analyzed with a qualitative analysis approach.

2.1 Selection of videos

YouTube videos were chosen as sources to be analyzed because one can receive the physicians' opinion directly and mostly unfiltered. For example, when reading an article where a journalist writes about physicians and AI, the journalist might (unconsciously) introduce a bias. These articles are rarely completely neutral and objective.

To enclose the source pool and to find the optimal and most suitable sources for answering the research question, four categories were built, consisting of 1) Language, 2) Year of publication, 3) Content and 4) Number of sources. Furthermore, a search strategy for YouTube was developed containing special key phrases.

For the first category the main inclusion criteria was that the videos were published in English which offers the possibility to get an insight into the opinion of physicians from various countries instead of restricting to one country. The second, crucial criteria is concerned with the year of publication. Since the state-of-the-art AI is very rapidly and constantly changing, videos published before 2020 are excluded. Moreover, as explained above, to include a video in the source pool it is necessary that each video contains at least one physician's opinion on AI in the medical sector. The videos can either be made by a physician or can contain a physician who is talking at a conference or is interviewed by someone else etc. To clarify, a video displaying a human with another profession than physician discussing about AI in medicine would not match the criteria of the categories and would be excluded. Category 4 restricts the compatible sources to the number of 20 videos, because of two reasons: Firstly, even after considering the former three criteria, there are still a lot of compatible sources. The other reason is that the duration of the videos is not limited, which means that even long videos are allowed. So, for a manageable research the number of sources had to be more restricted. Therefore, since all videos must be watched carefully and transcribed, it became clear that 20 videos are the maximum workload according to time and members of the research team. For the search strategy, four key phrases were created to search on YouTube for the most matching sources, namely:

1. How are doctors thinking about AI?
2. What do physicians think about AI (in medicine)?
3. AI in healthcare doctor's opinion
4. AI in medicine physicians' opinions

The sources were collected in a shared Google Sheets document until both research members watched all videos and agreed on the compatibility of each video. The analysis approach was based on Mayring, which will be explained in the following.

2.2 Qualitative analysis based on Mayring

To analyze the material in consideration of the research question, one of the following three independent analysis techniques (Mayring 2015) had to be decided on: Summary, Explication, or Structuring.

For the given research question, the technique *structuring* was chosen. With this method, the aim is to identify and to filter some aspects out of the material to evaluate the material based on predefined categories or criteria (Mayring 2015).

Additionally, Mayring (2015) distinguishes between four forms of structuring which differ in their actual aims, namely 1) formal structuring, 2) content structuring, 3) typifying structuring and 4) differentiated scaling structuring. Here, content structuring was chosen because this form indicates that certain topics, contents, or aspects of the material are to be filtered out and summarized with the help of categories. Therefore, the categories were built based on the ontology of the course and look like this:

1. Do the physicians have a positive, neutral or negative view of AI?
2. What area are the physician mainly talking about? / In what area are they specialized?
3. What do the physician say about losing their job because of AI? Are they afraid?
4. Are there future strengths/opportunities for AI in medicine/healthcare according to the physicians? If yes, which ones?
5. Are there any weaknesses/threats according to the physicians? If yes, which ones?

The analysis was done manually with another shared Google sheet document. Each category built one column and each source one row. The videos were watched focusing on the categories. When a statement of a physician matched one category, this statement was transcribed with the help of the YouTube transcript function that automatically displays a transcript of each video. Afterwards, the statement was inserted in the corresponding cell of the table. However, it must be noted that this transcript function is not 100 percent accurate, which is why some adjustments were done manually to display the exact words of the physicians. When one physician did not mention the topic of a category and did not make a compatible statement, the cell of the table was left out.

3 Results

Subsequently, each of the 20 videos was analyzed based on the five mentioned criteria. The first category is labeled as "physicians' view". 19 physicians have a promising attitude towards AI and feel entirely positive about the topic. However, one radiologist labeled AI as worrisome and had a rather negative view on this issue. The next category is labeled as "physicians main specialization". The results are depicted in Table 1.

Table 1: Frequencies of specialized areas

	Frequency
General medicine	7
Radiology	5
Cardiology	2
Diabetic eye disease	2
Cancer	1
Colonoscopy	1
Drug optimization	1
E-health	1

As one can see, the specialized areas that the physicians are working in are very different. However, not all physicians explicitly mentioned their field, they were just talking about AI in medicine in general.

When it comes to the public discourse about AI one considerable concern is that artificial intelligence might replace humans and thus threaten their jobs. Therefore, the third category is classified as "Job Loss". The analysis showed that, unlike the common beliefs, the majority of the physicians are not afraid of being replaced by AI and its related technologies and strongly agree that they will not lose their jobs due to AI. Instead, many physicians view AI as a tool which cannot replace them but assist them, as one physician explained:

> The thing is, digital transformation is often misunderstood as like a prime or core technology innovation and to some form it is but it is more about using the technology and including it into your daily practice as a physician, as a tool. Technology is not replacing physicians but it's augmenting their abilities like lab results or imaging has augmented our possibilities in the past. (of European Doctors 2020: 1:08:08)

Accordingly, they agree that AI can assist and support them to be more effective in the care that they provide and that AI can help to improve the diagnostic accuracy. One physician (Madai 2020: 2:34) went a step further and pointed out: ”AI will not replace medical doctors soon, definitely not do that. It will help them, so it’s the combination of artificial and natural intelligence that’s very important”. Some of them also mentioned the case that AI can accelerate the progress in medicine because it brings people from different occupational groups together and connects their knowledge. For example, as stated here (Peng 2020: 8:56), ”I believe if doctors, scientists and engineers work together we have the opportunity to address some of the biggest challenges in healthcare and help all of us live happier healthier lives.”

In contrast, another physician (Lungren & Lehmann 2020: 3:14) answered ”Yes, I hope so!”, when being asked if she thinks that she is training her replacement. So, she thinks that physicians will be replaced by AI technologies but is not worried or fearful about this change because she acknowledges the advantages of AI which she enumerates after the aforementioned statement.

Despite these positive statements, eight physicians did not mention the concern of losing their jobs at all.

The two final categories attend to the physician’s notion about weakness and threats as well as strengths and opportunities for AI in medicine. In the table 2, the latter are depicted.

The most frequently mentioned strength is the availability and accessibility of AI all around the world, that enables physicians to expand their impact because people worldwide can have access to healthcare, as stated by a physician (Parsa et al. 2020: 6:25), AI ”offers a means of access to primary care and by extension specialty care that wasn’t easily available before”. A lot of physicians share this opinion and additionally expressed their view on the benefits of having more access to healthcare all around the world. For example, a better access to medical services, realized by a stronger incorporation of AI in the healthcare system, also influences other factors like costs, as explained here (Abramoff & Frist 2020: 4:44), “It allows you to have access to very high quality care at very low costs anywhere where the patient is”.

The facilitated access can not only affect material factors like money. Sometimes people do not go and see a physician because their duties keep them from taking care of themselves. One physician (Peng 2020) added that this obstacle can be reduced with AI. She explained that especially people who suffer from certain diseases and live in rural regions normally have to travel very far to see a specialist who can help them. It often happens that they do not go and get the care they need because they do not find someone to care for their children while

Table 2: Frequencies of mentioned strengths of AI in medicine

	Frequency
Availability/access	7
Accurate diagnosis	6
Automated documentation	3
Better decision making	3
Faster diagnosis	3
Lower costs	3
Personalization of treatment	3
Dealing with big data	2
Time saving	2
Automated summarization	1
Doing routine tasks	1
Optimized treatment	1
Optimized work	1
Predict patients' need	1
Relentless	1

they getting their treatment or it is simply too expensive to go. However, she found an AI-based solution which offers greater access:

> So now with the AI installed in the facilities closer to where they live, patients can get care easily and efficiently and this means that they don't have to choose between caring for themselves and providing for their loved ones. (Peng 2020: 6:16)

Moreover, another often named opportunity is that AI provides a more accurate, earlier and faster diagnosis in comparison to human diagnosis, which helps especially the patients most efficiently. A physician (Goyen 2020: 2:25) explained that, ”the patients are the big winner in this game. AI will simply enable better diagnosis and earlier diagnosis with better treatment options so the patient is the winner”. The majority stated that a more accurate diagnosis can support them in their decision making, increase the quality of treatment and optimize their daily work. Additionally, an AI-based device can be so accurate, that it equals physicians regarding their diagnostic skills. “The algorithm we trained turned out to be pretty accurate and over the last few years we've been improving it such that now it's on par with retina specialists", as explained by a physician (Peng 2020: 4:48) who developed an AI tool to diagnose diabetic eye disease early and to prevent blindness. Another retina specialist went a step further and admitted that AI is already better than even experienced specialists, including him (Abramoff & Frist 2020).

However, the use cases of AI are not limited to a diagnostic tool. Instead, it can serve as a guidance tool which rather supports the physicians instead of providing a diagnosis itself. For example, some physicians explained that AI is able to support them in case they are not sure whether to treat a patient according to a specific diagnosis or not. Another one (Goyen 2020: 1:08) added that "AI can help the radiologist to highlight critical cases [...]. AI tells you 'Look at those images first because it's very likely that the diagnosis is there'".

Further important aspects that the healthcare sector can benefit from applying and utilizing AI are the opportunities of individualized treatments or time saving during the treatment process. For example, AI can cover routine tasks like paperwork which gives the physicians more time for the actual treatment. To go one step further, the physicians could be able to concentrate on patients with serious diseases if an AI treated minor problems by itself. A physician from California described that this can improve the patient-physician relationship:

> We can use AI to fix this problem. One is at the level of doctors. So, when we see patients instead of typing on a keyboard and looking at a screen, natural language processing can take that conversation and make a synthetic note. That's far better than any notes that we have today and also of course the ability to review all the data of a patient. That goes through many different levels, not just the electronic health record but the genomics and the sensors and all these different sources of data that weren't customarily available in any given patient. So, on that side we have harnessing AI tools to bring back that relationship. But on the patient side, it's getting rid of the need for a doctor for simple non serious matters like a urinary tract infection, a skin lesion or a child's ear infection. All those things today have been validated or are in the mists of getting validated for a doctor's diagnosis many times at levels of accuracy as good or even better than studies with doctors. So, we have an opportunity if we use both of these different pathways for AI support both for doctors and patients to bring the doctors and patients back together. (Topol 2020: 0:54)

Also, many of the physicians agreed that all these mentioned opportunities of AI will shape the clinical world and will cause a positive effect on medicine and the treatment of patients in general. For example, as stated here (Ho 2020a: 5:27), ”With AI we can now personalize and optimize treatment for each patient and it is my hope that we continue to push the boundaries of medicine with AI and save more lives in the future”. Another physician added:

> The last thing is that I really think that AI is going to make us much more efficient and we will be able to help more people in less or the same amount of time. You know what I mean, I think it will help bring access to places that don’t already have it. (Gupta 2020: 6:36)

Despite all these positive effects of AI in healthcare, also several weaknesses were mentioned. The results of the last category are shown in Table 3.

Table 3: Frequencies of mentioned weaknesses of AI in medicine

	Frequency
None	10
Demanding development	2
AI has limitations	2
Physicians need to learn how to handle the technology	2
Regulations/licenses determine work	2
Black box/no transparency	1
Challenging implementation	1
Concerns/worry of patients	1
Conservative clinics/physicians	1
Constant control needed	1
Makes mistakes	1
No one-size-fit approach	1
Patients’ background is missing	1
Question of liability	1
Rights on data	1

Two of the physicians labeled the development of AI-based techniques as tough, demanding and also time-consuming work. ”But training an accurate algorithm is really only the first step and there’s so much more work to do and we can’t do it alone”, informed one physician (Peng 2020: 4:58). The other one added that the technology must meet a lot of requirements:

> The hard part is how do you implement this? […] So it does need to be real time it, ideally should be hardware agnostic, should work with any

> scope manufacturer and you'd like the data that's coming out of that AI to integrate with your IT-system. [...] We need to make sure that autonomous AI is working on a fast majority of patients. (Karnes 2020: 8:32)

However, not only the development of AI technologies for the medical sector can be tough, a large and far-reaching challenge can also be the subsequent extensive application in the clinical world. To make that possible a lot of work has to be done beforehand, as explained by three physicians:

> [...] but to put it into purposeful medical practice every day, in every office, in every hospital, is a different challenge and that is very much connected to kind of a national or European strategy to really introduce it in a broad context. (of European Doctors 2020: 1:14:16)

Another one added that a lot of licenses are needed before using new medical devices on a daily basis, which can be challenging to receive, especially for autonomous tools (Feldmann 2020). Moreover, even when these obstacles are overcome, there are still patients and clinics who are conservative about AI and their personal data, therefore not willing to apply and get treated by such techniques (Madai 2020). An additionally discussed weakness is the inability of AI to perform exactly like a human physician does. This challenges the clinical staff because they have to understand this inability to properly use AI-bases techniques: The physician clarified:

> [...] and it's very clear that at the same time it has limitations for example creativity, reacting to unpredictable situations, problem solving, critical thinking. This is not something that artificial intelligence can do and this is something that is reserved for humans. So, at the end of the day, what is really crucial is for doctors to understand the limitations of AI to understand how best they can use artificial intelligence. (of European Doctors 2020: 38:18)

He added that a new challenge lies within the education of the physicians as they need to get familiar with the technology in order to control and supervise it. Therefore, he believes that AI will affect physicians' prime duty:

> And so it's a question of literacy of understanding because the task of the doctor will be not only to control what the algorithm is doing but also then to be able to explain to the patient how the artificial intelligence has worked and what are the conclusions and to assume the responsibility for the diagnostic or for the medical conclusions, for the decision-making in medical conditions. (of European Doctors 2020: 42:52)

However, despite all these discussed challenges, the majority of the physicians did not mention any weaknesses of AI in medicine and only expressed positive effects.

4 Discussion

This research focused on physicians' perception of artificial intelligence in the healthcare sector. Therefore, 20 YouTube Videos of physicians were analyzed based on five content-based categories, which split the physicians' perception into different subtopics to get a more detailed view on their understanding of AI.

The first category pointed out that almost all physicians had a general positive attitude toward AI. Even so, some of them mentioned a few weaknesses or challenges which should be given more attention to in the future.

The second category highlighted the different areas the physicians are working in. It is worth noting that, although these areas are varied and dissimilar, many AI technologies are already applied within all these fields. This finding shows that AI can be utilized in many different fields of medicine and that it can support e.g. a radiologist as well as a cardiologist. This highlights especially the various potential use cases of AI in medicine, meaning that AI is not restricted to only one subfield of medicine. In addition, the training of a new AI-based device can bring scientists as well as IT-specialists and physicians together because it needs the support of different specialists. For example, when AI is trained to make a diagnosis based on medical images, it needs the knowledge of a radiologist to scan the image correctly. Nevertheless, it also requires other specialists like cardiologists or neurologists to interpret the image and detect a disease correctly. Furthermore, the algorithm itself can only be implemented by an IT-engineer, so all of these occupational groups might have to work hand in hand.

When looking at the third category about job loss, the results clearly match with the general positive attitude of the physicians toward AI. With 13 out of 20 physicians, the majority disagrees with the claim that they will lose their jobs because of AI. Also, most of them see AI as a chance which can support and simplify their work as a diagnostic tool like an MRI or lab results. The physicians noticed that even though AI could alter their daily practice, they are sure that it is not yet able to replace them. Furthermore, they mostly welcomed the application because they acknowledged the advantages for them as physicians, for medicine in general and especially for the patients. This sentiment can also be found when considering the opportunities and strengths of AI which were identified and mentioned by the physicians throughout the videos. It is noticeable that every physician found some advantages of AI in medicine. Moreover, as seen in category two, the mentioned benefits cover different subfields of medicine which shows that the use cases for AI are unlimited. The tasks range from individual drug optimization for patients, to supporting radiologists in interpreting images and saving time with documented conversation by natural language processing. However, the physicians agreed that AI can bring the same improvements no irrespective of the exact area of application. This is also reflected in the fact that benefits like a faster and more accurate diagnosis, time saving, lower costs and personalization of treatment were mentioned frequently and explained in detail by a lot of physicians. This shows that physicians acknowledge the strengths of AI because they notice the improvements AI brings to various subfields of medicine. As these findings correspond with those regarding category one and three, it can be speculated that the physicians do not fear the loss of their jobs because they are aware of various opportunities that comes with the application of AI technologies in their field.

The last category covered the challenges and weaknesses of AI in medicine. In contrast to the strengths of AI in category four, it stands out that weaknesses were mentioned significantly less frequent by the physicians. Half of them did not mention any weakness at all and only addressed the positive sides of AI. So for the physicians from the analyzed YouTube videos, the advantages that were raised definitely outweigh the mentioned challenges. This, again, highlights the physicians general positive attitude. Furthermore, it is important to mention that all stated weaknesses are solvable. For example, a mentioned challenge for the physicians is the application in every hospital and dealing with conservative patients, physicians and clinics. Naturally, this challenge is justified but with educational talks that demonstrate how advantageous and safe an AI device is to the physicians and clinical staff, it is possible to assure them of the benefits that comes with the application of AI. In addition to that, educational talks might also help to reduce the worries and concerns of patients e.g. fairness of the system or the risk of harm which was mentioned by another physician (Abramoff & Frist 2020). Another often raised issue is that companies that develop medical AI devices have to get licences, which is a big obstacle that needs to be overcome before AI can be utilized in the healthcare sector. Although this challenge can be tedious, an assumption is that companies might get licenses faster the more AI-based medical devices are developed and the further the research has progressed. Overall, one can see that the mentioned weaknesses are all solveable but some with more effort than the others. This indicates that the physicians acknowledge the new chances AI is bringing the medical sector but also draw attention to the weaknesses which have to be solved before AI can really unfold its actual potential. Therefore, these challenges should still be taken into account and discussed because they have the potential to improve the technologies by revealing difficulties and ethical problems. Again, these findings of the last category coincide with the other categories because the few mentioned weaknesses underline again the general positive attitude of the physicians toward AI.

5 Conclusion

5.1 Implications

In conclusion, the analysis based on the five categories shows that the physicians from the source pool generally have a positive attitude toward AI in medicine, which can be seen in the facts that they 1) do not fear a job loss due to AI, 2) find considerably more strengths than weaknesses and 3) most weaknesses are solvable and offer the potential to improve the technology after solving. Therefore, this research indicates that physicians are quite welcoming toward artificial intelligence because only a decent number of physicians expressed worries or weaknesses at all. For them, constant availability of AI and equality of medical care for all people around the world is the most important and the most obvious opportunity, followed by a faster and more accurate diagnosis due to the use of AI. The few declared weaknesses covered issues like high workload, regulation by licenses and general concerns by conservative patients as well as clinics.

5.2 Limitations and outlook

The number of 20 physicians offers a brief insight in the general opinion of physicians towards AI. However, this is not adequate enough for generalizing over all physicians which restricts the given findings. Furthermore, the field of AI in the medical area is not discussed exclusively in English. Therefore, in order to have a more precise research, including assumptions of physicians who speak other languages than English is decisive to get a deeper insight.

To pursue this research, an interesting topic for future studies is the view of other clinical members about AI because the health care sector is not only limited to physicians but also includes clinical practitioners, nurses and medical assistants. Therefore, to get an insight into how AI is generally perceived and discussed in the health care sector, future studies have to investigate different occupational groups. This might offer the opportunity to compare the opinions of different professions and to see if for example nurses perceive AI differently than physicians. However, the findings of this study provide a reasonable starting point for further investigation.

To conclude, further research should repeat the investigation with a larger population of physicians and people from other medical professions. It is crucial to open the search criteria and also include other languages as well as other sources of media i.e. newspaper articles, documentaries or podcasts where the clinical staff can express their opinions.

References

Abramoff, Michael & Bill Frist. 2020. *Michael Abramoff, MD, PhD: Founder of IDx on artificial intelligence in medicine.* https://www.youtube.com/watch?v=8uGNRX2wRwE. (Accessed on 03/28/2021).

Bathaee, Yavar. 2017. The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law and Technology.* 31(2). 889.

Buch, Varun, Irfan Ahmed & Mahiben Maruthappu. 2018. Artificial intelligence in medicine: current trends and future possibilities. *British Journal of General Practice* 68(668). 143–144.

Currie, Geoff, Elizabeth Hawk & Eric Rohren. 2020. Ethical principles for the application of artificial intelligence (AI) in nuclear medicine. *European Journal of Nuclear Medicine and Molecular Imaging* 47(4). 748–752.

DePuey, Gordon, Ernest Garcia & Norberto Ezquerra. 1989. Three-dimensional techniques and artificial intelligence in thallium-201 cardiac imaging. *American Journal of Roentgenology* 152(6). 1161–1168.

Dilsizian, Steven & Eliot Siegel. 2014. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current cardiology reports* 16(1). 441.

Feldmann, David. 2020. *What are the key telehealth risks facing physicians?* https://www.youtube.com/watch?v=yCicCnFzLec. (Accessed on 03/28/2021).

Goyen, Mathias. 2020. *How AI will help doctors and ultimately patients: Prof. Dr. med. Mathias Goye.* https://www.youtube.com/watch?v=P8eI-uodk38. (Accessed on 03/28/2021).

Gupta, Yasha. 2020. *AI in radiology | Yasha Gupta, MD.* https://www.youtube.com/watch?v=lh54RCKAiBE. (Accessed on 03/28/2021).

Ho, Dean. 2020a. *AI gives muscle to medical research.* https://www.youtube.com/watch?v=-NQG3kQrh40. (Accessed on 03/28/2021).

Ho, Dean. 2020b. Artificial intelligence in cancer therapy. *Science* 367(6481). 982–983.

Karnes, William. 2020. *Artificial intelligence in GI | William E. Karnes, MD, UCI | ULCA Health Digestive Diseases*. https://www.youtube.com/watch?v=rZ31xIVoLDk. (Accessed on 03/28/2021).

Khoroshevsky, Vladimir, Vladimir Efimenko & Irina Efimenko. 2020. Artificial intelligence, biotechnology and medicine: reality, myths and trends. In *Russian conference on artificial intelligence*, 416–436. Moscow: Springer.

Lungren, Matthew & Constance Lehmann. 2020. *This job is being replaced by an algorithm*. https://www.youtube.com/watch?v=F8A30CL2GHs. (Accessed on 03/28/2021).

Madai, Vince. 2020. *How artificial intelligence is augmenting medical doctors - I am a doctor and AI coder*. https://www.youtube.com/watch?v=rEOkabbnxB4. (Accessed on 03/28/2021).

Mayring, Philipp. 2015. *Qualitative Inhaltsanalyse*. Weinheim: Beltz Verlagsgruppe.

of European Doctors, Standing Committee. 2020. *Doctors going digital: how to future-proof skills*. https://www.youtube.com/watch?v=XEdf6CN1Kmg. (Accessed on 03/28/2021).

Parsa, Ali, Mobasher Butt & Robert Fields. 2020. *How Babylon is using AI to revolutionize access to healthcare globally*. https://www.youtube.com/watch?v=218lD-c3RgQ. (Accessed on 03/28/2021).

Patel, Darshan, Yash Shah, Nisarg Thakkar, Kush Shah & Manan Shah. 2020. Implementation of artificial intelligence techniques for cancer detection. *Augmented Human Research* 5(1). 1–10.

Peng, Lily. 2020. *Democratizing healthcare with AI | Lily Peng | TEDxGateway*. https://www.youtube.com/watch?v=MNp26DgKxOA. (Accessed on 03/28/2021).

Topol, Eric. 2020. *Can AI improve doctor-patient relationships?* https://www.youtube.com/watch?v=uyQ-XWsgtl8. (Accessed on 03/28/2021).

Watanabe, Alyssa, Vivian Lim, Hoanh Vu, Richard Chim, Eric Weise, Jenna Liu, William Bradley & Christopher Comstock. 2019. Improved cancer detection using artificial intelligence: a retrospective evaluation of missed cancers on mammography. *Journal of digital imaging* 32(4). 625–637.

Part III

The media

Chapter 12

How AI transforms public discourse — An analysis of the impact of AI in public discourse as portrayed in major news outlets

Nikolai Godt, Ivan Polivanov, Karina Khokhlova & Muhammad Faraz Rajput

As Artificial Intelligence (AI) technologies become more visible in our daily lives, their impact is regularly reported in social media. Since this book explores the understanding of such a phenomenon as AI in Public Discourse (PD), we provide a study of the impact of such technologies, namely: Generated Media and Personalization Algorithms on public discourse from 2018 to 2020 among the four most popular English-written online newspapers. This paper demonstrates how media coverage can stimulate public discourse about new technologies. Examining the benefits and challenges of AI being portrayed in the media can lead to a deeper understanding of the potential ramifications of the public discourse, especially in relation to the development and influence of AI technologies. In this work, we used such text analytical and annotation systems as: *CATMA* and *IBM Tone Analyzer*. Our results show that the media have a fairly realistic and practical focus in their coverage of the impact of AI.

Keywords: Public Discourse | Deepfake | Personalization Algorithm | Social Media

1 Introduction

1.1 Artificial Intelligence

Over the past few years, applications in the field of artificial intelligence reached new heights in popularity(Yang 2018). Nowadays, almost all areas of (human) life, from business contexts to use-cases in medicine. AI can successfully play complex games against humans and even write code on its own. While this rapid development has not yet been acknowledged by society to its full extent, its implications and consequences are barely in the awareness of society. Eventually, some technologies remain oblique to the users but play a significant role in the perception of information. In recent years, a series of political and societal events, such as elections, can be in part traced back to the technology of AI, which causes concerns among scientists and experts(Chattopadhyay 2020).
There are many uses of AI to explore, however, in this paper, we focus on AI as Algorithms that can

1. select media and information for an end-user, and
2. create media artificially

1.2 Personalization Algorithms

Currently, we live in a period referred to as the information age. The amount of data we produce and that we are theoretically able to access is increasing repeatedly each year. For organizations requiring this data, it gets more and more difficult to deal with this load of information. A solution to this problem is offered by ”Personalization Algorithms”, also called ”Recommender Systems”. Based on previous data, they automatically e valuate the relevance or importance of new data and can therefore prioritize the information that the algorithm deems most important. This prioritizing, however, is different with every user, depending on his or her interests or online behavior.(Gupta 2013)

In September 2011, Facebook changed users’ news feeds by introducing a new machine learning algorithm. Prior to that, users used to see posts in their feed in chronological order, with the most recent posts on top. After the change, the users were presented with a feed that was predicted by the algorithm to be most likely to be interacted with by the user.

Many other companies started to make use of personalization to improve customer satisfaction, increase digital sales conversions, marketing results, branding and improve website performance, and for advertising. Personalization became a key element in social media and the internet as a whole.

Seven years later, in March of 2018, the Cambridge Analytica Scandal, by an article in the New York Times(Rosenberg et al. 2018), exposed what an enormous influence the Facebook recommender system had on the 2016 American election and how its algorithm can be used for manipulating and distorting public discourse. The recommender systems on Facebook were used in Donald Trump's election campaign to specifically target the people that seemed the easiest to manipulate with election advertising. After this scandal, Facebook made some minor changes to the algorithm but ultimately, the way how posts were shown to the users did not change significantly.

On a more general note, the perception of reality as a whole can change remarkably when constantly exposed to the content streams that follow a specific pattern largely with the most sensationalist and emotionally loaded content, made possible by these Recommender Systems. The authors of chapter 25 in this book 16 researched the consequences of social media for society and individuals with the example of "The Social Dilemma" and therefore is in a quite similar territory as our one field of study in this paper.

1.3 Generated Media

"Synthetic Media" or "Generated Media" describes media that is created purely algorithmically and automatically. It is important to note that, while sometimes used in public discourse as synonymous to Synthetic Media or Generated Media, the term "Deepfakes" actually describes a specific technique of synthetic media that has been largely made possible through the developments of General Adversarial Networks (GAN) in the recent years. Deepfake technology analyzes real images or videos from human faces, depicting various facial expressions, and creates a video that can map the facial features of a face so well onto another human, that it is quite difficult to spot that the face actually does not belong to the rest of the body. The GAN, which is generating the video, consists of two AI algorithms, called the generator and the discriminator. The generator is creating an output that should depict an object or person, in the case of Deepfakes a human face. The discriminator tries to detect a face in the output and feeds the evaluation back into the generator. With this information, the generator can iteratively improve the quality of the output until a certain threshold is met and the generation process is finished.(Goodfellow et al. 2014)

Next to Deepfakes, there are also other synthetic media like natural text generation, which recently made a huge leap forward with the development of GPT-3 an autoregressive language model that uses deep learning to produce human-like text(Brown et al. 2020).

In this paper, we will only work with the term “Deepfake” but we class it as a representative example for synthetic media as a whole. The first time that this Deepfakes gained widespread public attention was through an investigative article published by Vice(Cole 2017) in November 2017. The report shed light on the use of Deepfakes for porn videos, uploaded onto a specific Reddit page. Users uploaded their own videos made with the Deepfake algorithms that placed faces of popular actresses onto the bodies of porn actresses in porn scenes. As a result of the article and the following discussion, Reddit banned the subreddit and all use of Deepfakes on its platform.

This way of dealing with the problem is not uncontroversial in itself, as it imposes a form of censoring into the public discourse, even if its intentions are only positive in nature. Please refer to Chapter 25 of this book, as it includes extensive analysis on ”Content Filtering”. Their research critically examines one way of dealing with the supposedly problematic consequences of “fake” or “synthetic” media, which we will, next to other measures, investigate in this paper as well.

1.4 Motivation

The quality of a public discourse depends on the people who participate in it and the information that the participants have about the matter of discourse.

Using automated tools to decide how information will be distributed to the people, has very likely direct consequences for public discourse. AI that is capable of creating media which looks convincingly real but actually is not, will also likely shape public discourse, as participants of the discourse might falsely use this media as arguments in a discourse.

With the two articles from The New York Times and Vice that brought the two algorithms into the light of the public, we have a quite negative outlook on the technologies right from the start. That motivated us to analyse all consequences and possible actions for the technologies as it is displayed by those influential and widely read newspapers.

Despite the apparent overwhelming downsides of these technologies, we also conducted a sentiment analysis to get a detailed metric of emotional and linguistic tones in the articles.

Public discourse as a whole is a complex topic with many different agents and places that shape and build the discussions. Newspaper outlets have, as the two above mentioned articles show, a significant influence on the public discourse about AI itself. They reach a bigger and broader audience than science journals and they have to make the sometimes complex topics they report on understandable to the reader while keeping the matter accurate. They have the role to inform the general public and therefore are a good indicator on what is discussed in public.

That is why we in the following present a brief overview over the perception of major newspapers on the before mentioned technologies.

2 Methodology

2.1 Source Gathering

To limit the number of articles we work on, we narrowed down our scope to a fixed temporal window, as well as a language and popularity window. All our articles are not older than from the 1st of January 2018. For reasons of comparability and feasibility when working together as an international team, they are exclusively English speaking, Not all newspapers have the same audience reach, and since we want to get an insight into the "mainstream" perspective towards the relevant technologies, we picked the most popular news outlets worldwide, determined by web ranking(Media 2019). In total we selected the 4 most popular/ most read outlets in English language, namely: *The New York Times*, *The Guardian*, *The Washington Post* and *Daily Mail*.

From the newspapers, we retrieved three articles that talk about generated media (see table 1) and personalization algorithms (see table 2)each. We used the Google Search Algorithm with certain keywords to ensure that the articles are about the technology and also relevant. As keywords we chose *Deepfake* for Synthetic Media and *((Personalization "Algorithm") OR (Relevance AND "Feed" AND "Algorithm"))* for Personalization Algorithms. The searches usually yielded an extensive amount of articles, from which we always chose the first three for our study. Theoretically, we would have a total of 4 newspapers * 2 keywords for each newspaper * 3 articles for each topic = 24 articles. However, we only retrieved 22 articles, since some searches did not yield enough articles.

2.2 Research Questions

In order to conduct the analysis, we focused our attention on the following research questions regarding both technologies:

1. What kind of attitude do the articles have towards the technologies?
2. What companies are the most associated or mentioned?
3. In what way are society and public discourse changed and influenced?

Table 1: Selected articles for “Generated media”

The Washington Post
”‘Deepfakes’ are here. These deceptive videos erode trust in all news media” (Vaccari & Chadwick 2020) ”White House shares doctored video to support punishment of journalist Jim Acosta” (Harwell 2018) ”Top AI researchers race to detect ‘deepfake’ videos: ‘We are outgunned’” (Edwards & Livingston 2018)
The New York Times
”Internet Companies Prepare to Fight the ‘Deepfake’ Future” (Metz 2019) ”Episode 21: ‘Fake Believe’” (Schmidt 2019) ”Facebook Says It Will Ban ‘Deepfakes’” (McCabe & Alba 2020)
Daily Mail
”Elvis back from the dead? Artificial intelligence is used to create eerie ’deepfake’ pop songs that sound like they are being sung by deceased stars” (Chadwick 2020) ”Channel 4 will use a ’deepfake’ version of The Queen to deliver their version of her Christmas message...with jokes about Harry and Meghan and Prince Andrew” (Sharples 2020) ”Deepfake detection: Microsoft unveils a tool that can tell if a video or photo has been doctored in a bid to combat disinformation online” (Morrison 2020)
The Guardian
”What do we do about deepfake video?” (Chivers 2019) ”The rise of the deepfake and the threat to democracy” (Parkin 2019) ”To fix the problem of deepfakes we must treat the cause, not the symptoms” (Beard 2019)

4. Which spheres of public discourse are most affected?

5. What steps have already been taken and what demands of actions are voiced?

2.3 Text Annotation

Before we started annotating the text to get any statistically valuable information out of the data, we first went through a process of selecting the right tool to do so. There were multiple text annotation programs we considered, including *Sketchengine*, *Catma*, *TagTog* and *Inception*; finally we decided to go with *Catma*

Table 2: Selected articles for “Personalization algorithms”

The Washington Post
”A face-scanning algorithm increasingly decides whether you deserve the job” (Harwell 2019) ”The personal stylists who are training the bots to be personal stylists” (Bhattarai 2018) ”Dear tech companies, I don’t want to see pregnancy ads after my child was stillborn” (Brockell 2018)
The New York Times
”Personalization Has Failed Us” (Klosowski 2019) ”Is TikTok a Good Buy? It Depends on What’s Included” (Roose 2020)
Daily Mail
”Tinder for TV: AI swiping app can recommend shows based on what you like and what your friends are watching” (Liberatore 2018) ”Snapchat’s makeover arrives: App rolls out redesign that separates ’Friends’ and ’Discover’ pages to create a more personalised experience” (Pinkstone 2018)
The Guardian
”From viral conspiracies to exam fiascos, algorithms come with serious side effects” (Naughton 2020a) ”How Amazon puts misinformation at the top of your reading list” (Naughton 2020b) ”Facebook’s news feed change won’t help social media addiction” (Stefanou 2018)

since it is the only tool that allows for whole paragraph annotation, has a collaborative environment and offers analysis tools.

Before loading the articles into *Catma*, we copied them from the website into a .txt file to preprocess. We deleted all text that was not part of the article. In *Catma*, we created two projects, one for Generated Media and one for Personalization Algorithms in order to keep the data apart. Each article was then annotated with specific tagsets, which are tied to the reseach questions stated above. The tagsets are: “Companies”, “Affected Places of Public Discourse”, “Types of Influence” and “(Demands of) Actions”. We based our tags on the ontology from the course and modified the ontology wherever necessary in order to fit the topics. Depending on the object of study, the internal tags of these tagsets have different definitions.

2.4 Analysis Procedure

To analyze the annotations, we used the *Catma* Analyze Tool which can filter by tag or keyword. For each tag, we gathered both the percentage of the tag in relation to the whole set of tags in a tagset, as well as the percentage of articles this tag appeared in. This is an important distinction since a tag can occur a lot in one article which specifically talks about that tag, while no other article has the tag in them. The tagset percentage would then be quite high while the article percentage would be low.

In addition to *Catma*, we relied on *IBM Watson Tone Analyzer*(*IBM Watson*) to provide us with sentiment analysis of our data. For that, we loaded the .txt files into IBM Watson and assigned those sentiment tags to each article, which the analyzer had >50% confidence on. Our sentiment analysis contains two classes: Emotional Tones and Language Tones. There are 7 tones in total. For the first class - Anger, Fear, Joy, Sadness. For second - Analytical, Confident, Tentative.

3 Results

For our work, we have created a dataset of 636 tags: 331 for Generated media and 305 for Personalized Algorithms. They were sorted by the categories selected in the defined ontology. In this section, we represent the collected data according to their categories and conclusions about their impact on the course of our research.

3.1 Personalization Algorithms

3.1.1 Sentiment

From the sentiment analysis with IBM Watson, a majority of articles (90%) were classified as containing a joyful tone. Half of the articles contained a significant amount of text that was classified as sad. 10% of the articles included a fearful tone while no article was reported to contain anger. Many articles contained both a “Joy” and a “Sadness” tag. This result in itself is not contradictory since IBM Watson assigns the tones for each paragraph and then classifies the whole article with those tones that were sufficiently often assigned in the paragraphs.

All articles contained an analytical tone and 90% a tentative tone as well. There was no article with a tone of certainty.

3.1.2 Tech Companies

During the annotation stage, twenty-three company/platform names were identified. After the analysis of the number of mentions of all the collected companies and platforms, it is possible to identify the most influential one for the specified topic to answer the first research question. Results show that TikTok with its amount of mentions takes 26.67% among all denominations. TikTok has the highest value in this category. However, TikTok was only mentioned in one article which does not meet the criteria established in the methodology of this study. For these reasons, Facebook is the most influential company in the field of Personalization Algorithms. Facebook has the highest ratio for both analysis requirements: for mentions – 17.19% and for references by different articles – 75%.

Moreover, the presence of relationships among companies and platforms should also be highlighted. Since Facebook owns the social network Instagram, and a messaging app WhatsApp, which also have 4.62% and 0.51%, respectively, it only underlines the influence of Facebook on public discussions of AI technologies in the press.

3.1.3 Affected Areas

The places that are most influenced by Personalization Algorithms are Social Media, Business, and Culture. They have 31.82%, 25.00%, 20.45% of the whole data tags respectively, as shown in figure 1.

If we review the result in terms of mentions through the articles, we can observe a completely different picture. In the first place among the mentions in articles is Social Media area- it is mentioned in 7 articles out of 10. And then Business, Culture, and the Individuals in Society go on equal positions - they are all mentioned in 5 articles out of 10 (fig. 2).

Compared to Generated Media, which will be described later, there is no tag “Politics” throughout the whole amount of articles.

3.1.4 Effects

Compared to Deepfakes, the classification of the tags in Personalization Algorithms in positive and negative was way more balanced. Of all tags, about 53% were negative and 47% were positive.

As depicted in figure 3, the by far most described effect out of all was that these algorithms save the user most of the time with deciding what information or media to engage with. This described Convenience also can lead to increased efficiency especially in the business environment. It saves the HR departments

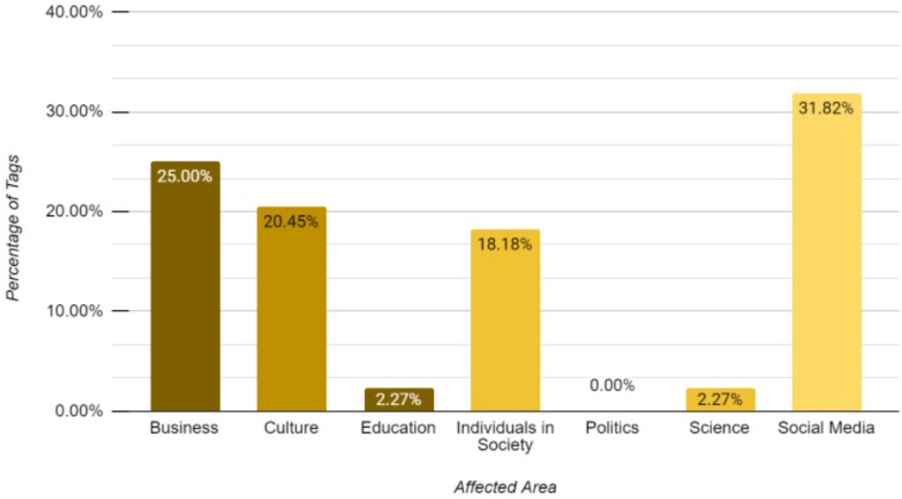


Figure 1: Affected Areas for Personalization Algorithms

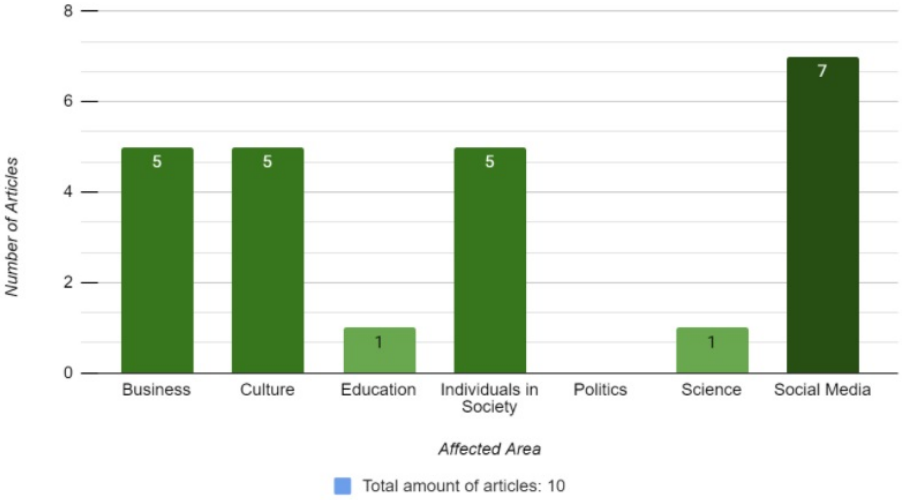


Figure 2: Affected Areas for Personalization Algorithms

significant time when searching for new employees as the algorithm can – even before the interview – rank the appliers by their evaluation fitting to the companies expectations. We categorized this effect under the aspect of the Improving Economy. Most of the time, the recommender systems are deployed not to give the best content to the users, but rather to optimize the advertising scheme so that the companies find an easy way to reach their target audience, which can be determined to a very detailed degree.

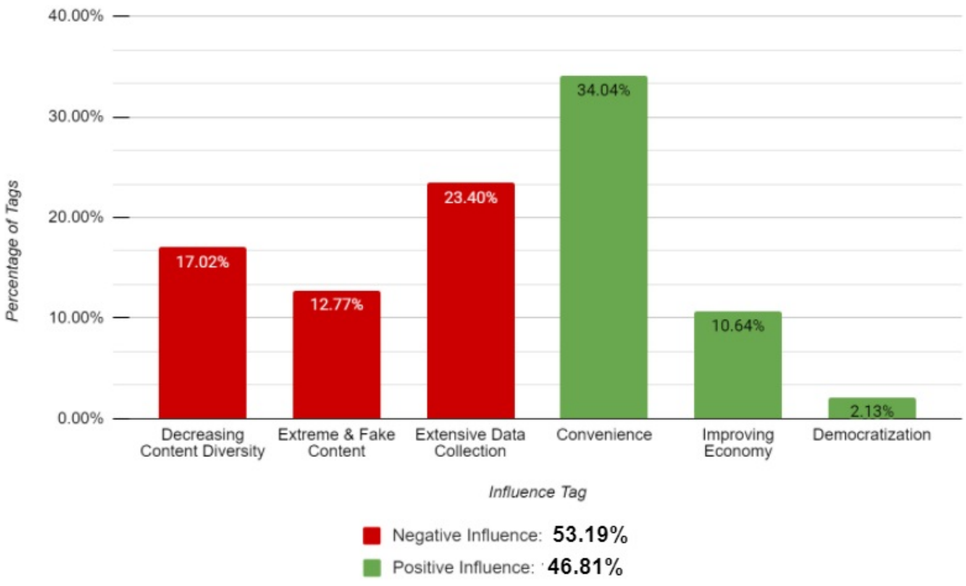


Figure 3: Influence Evaluation for Personalization Algorithms

The third of the positive aspect, through only associated with ca. 3% of the Tags, can be described as the democratization of content and the respective diversification of content producer to contend consumer relationships. A good example of that is the music industry, which through the advent of streaming platforms like Spotify, excessively rely on those Personalization Algorithms, has been diversified to a great extent. While even just a decade ago the majority of the music was released through the means of a few influential music labels, nowadays through the connectedness it is way easier for independent artists who serve a niche to find their audience.

3.1.5 (Demand of) Actions

As the methodology shows, to define what kind of action was mentioned the most, we used our tags among 12 articles. Collected data according to their number of appearances was compared and it was found that Moderation has the highest result around 59%. Which is more than half of all actions against AI discussed in these articles.

It can also be seen from the collected data illustrated in Figure 4 that education, in addition to moderation, also prevails with 21%. Finally, prohibition and government/law have the same values and the lowest among the indicated actions, with 10% each (fog. 4).

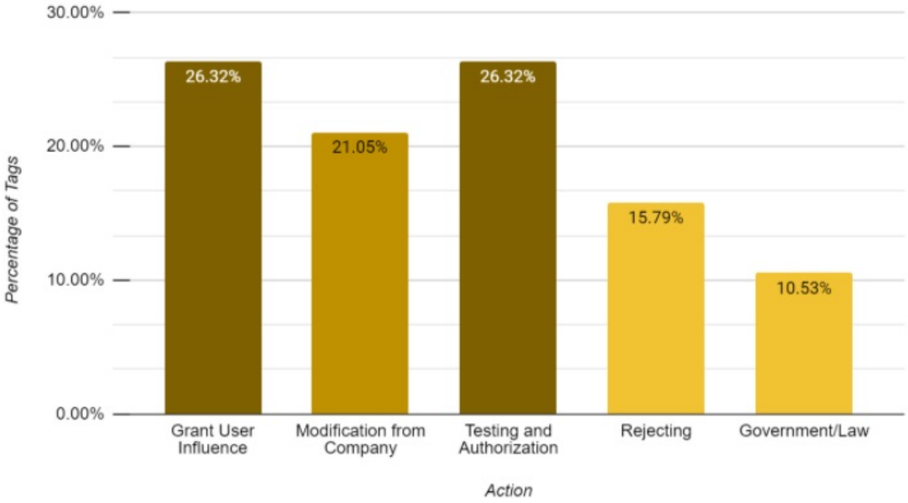


Figure 4: (Demand of) Actions for Personalization Algorithms

It is important to note that the aforementioned results went depending on the number of mentions among all articles, however, if we pay attention to the number of mentioned articles for each of the tags, it can be seen that governmental influence was still discussed more often, which indicates more significance of this side of action in comparison with prohibition itself.

3.2 Generated Media

3.2.1 Sentiment

Similar to the articles about Personalization Algorithms the articles for Generated Media mainly contained emotional tones of joy and sadness, while fear and anger were not many presents. In this case, however, sadness is contained in 83% of all articles and outweighs joy which was classified in 75% of all articles. Fear was detected in 17% of all articles and anger again was found in none of the articles.

In regards to the language tones, IBM found 92% to be analytical and 83% tentative while no article was classified as containing a tone of certainty.

3.2.2 Tech Companies

Results show that Facebook with its number of tags takes 31.13% and the Number of articles with 75%, is the highest among all columns that shows how many times the companies had been discussed and the name of the articles as well. Facebook is popular among all social media platforms. Seventeen billion users visit a Facebook page in a year, making it the fastest way for spreading any kind of information.

3.2.3 Affected Areas

The places that are most influenced by such technologies are politics, social media, and individuals in society. They have - 33.7%, 28.26%, 16.30%, respectively, as presented in figure 5.

Reviewed the result in terms of mentions through the articles, it was observed that the overall result did not change. The main areas are still politics, social networks, and individuals in society. The only difference is a subtle distinction between the percentage. It is 83.33%, 66.67%, and 58.33%, respectively. The reason for such difference is the number of articles containing tags mentions, they differ from each other by only 1-2 articles. Thus, for example, politics is present through 10 articles out of 12; social media – 8 out of 12; and individuals in society – 7 out of 12 (fig. 6).

On the other hand, the smallest number of tags can be observed in business and science. They have 6.52%, 3.26%, respectively. Both tags are extremely rare and are met only among 2 of the 12 articles, as depicted in figure 5 and figure 6.

Also, a tag representing an area of education is absent through all datasets of articles.

3.2.4 Effects

When analyzing the different effects that Generated Media have on Public Discourse a clear tendency towards negative effects becomes apparent. 87% of all

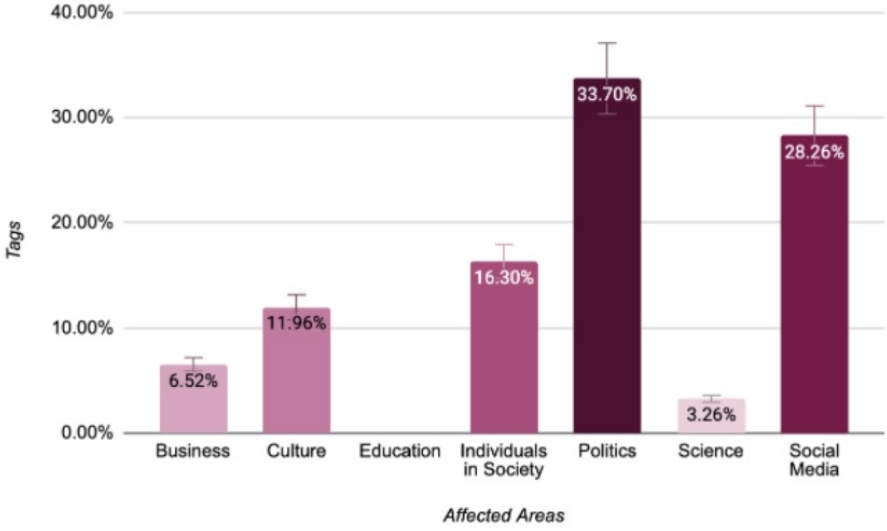


Figure 5: Affected Areas for Generated Media

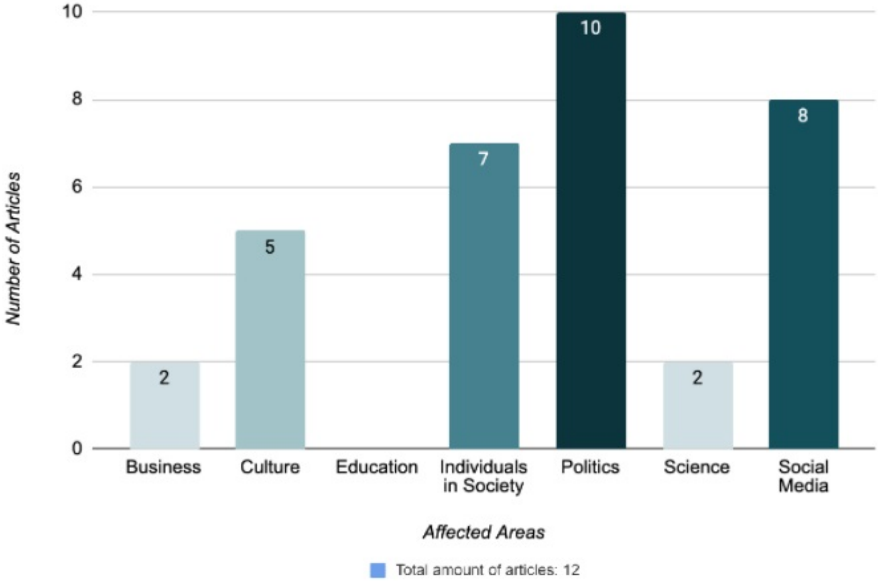


Figure 6: Affected Areas for Generated Media

Tags from the Effect Tagset we have classified as negative compared to only 13% Positive Effects mentioned in the Articles.

A closer look at the individual Tags reveals that the possibility of Manipulation was by far the most mentioned Consequence of the deployment of Generated Media Technologies (34% of all Tags). That is closely tied to the democratic vulnerability that became especially apparent in recent years, in which the possibility of influencing foreign elections with the help of disinformation has possibly been used. This way of exerting power can obviously be also used with synthetic Media as it holds even more opportunities to twist the perception of the masses to create a certain political outcome (Fig. 7).

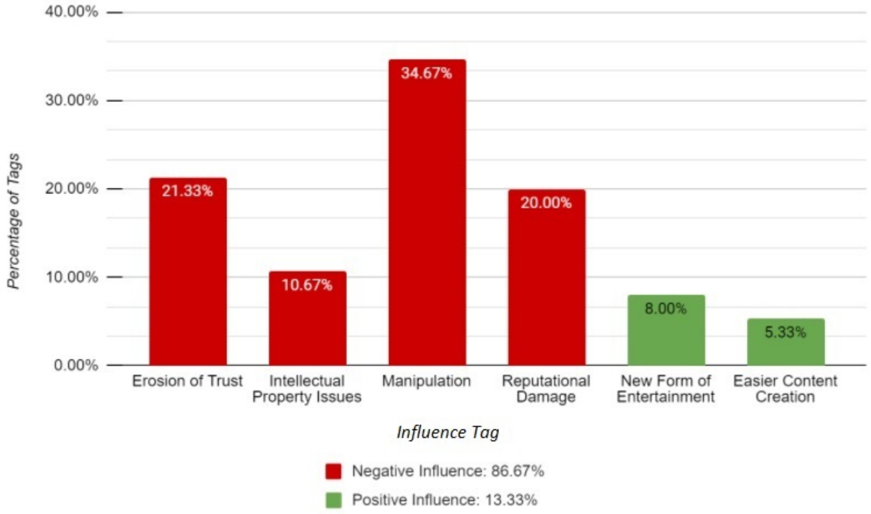


Figure 7: Influence Evaluation for Personalization Algorithms

Accompanying that, we found that roughly one 5th of the tags have been associated with the consequence of the erosion of trust. If not the obvious power of Deepfakes is not enough to create a sense of insecurity in what to trust, then, at last, the effects of those previously mentioned misinformation campaigns will create a sense of distrust in society. While The Guardian actually found this fact a positive thing, since it makes people more aware to not trust every single thing that can be seen on the internet(Parkin 2019), all other authors classify this as a negative thing, hence the Tag is included in the ”negative” Tagset.

Next to the consequences that largely affect the whole society and their political systems, the effect of reputational damage also, if not mainly, affects people more on an individual level. The creation of deepfakes as a form of revenge or bullying has already been reported several times. The victims of these are attacks, who are usually depicted in the Deepfakes in some form, can suffer from social alienation, depression, and anxiety.

The last form of reported negative effects can be described as intellectual property issues. With the creation of synthetic media, a dataset is needed to feed the algorithm in order to output some media. This dataset usually includes media created by people, which the AI is supposed to mimic. Since the created output only mimics the input, (synthesizes) the question of who owns the created piece of media remains debated. Mostly this issue was reported in connection with the music industry. In an article from the New York Times, they reported on a project in which the composed music of Elvis Presley has been synthesized in order to create other pieces that resembled the originals in form and style. Whether that new music is just "inspired" or actually "copied" from the original artist is a matter of perspective on the process of creation.(Chadwick 2020)

The positive sides of this technology have sorely been the improved ease of content production (5%) and the possibility of creating new forms of entertainment(8%). To deploy an algorithm that can create humans in a professional look and sets in mere seconds, is an extremely efficient tool for advertising companies as they can save on the cost of models and photographers who would have been paid in the production process.

3.2.5 Demands of Action

For the "Demand of actions" tag set in research of this side of AI technologies, we used different tags as was presented in our methodology. From a number of mentions of those actions groups we have a result but not as clear as it was with Generated Media. In this case, no action above 30% was found.

However, "Grant User Influence" and "Testing and Authorization" have the same amount of 26% which can be called most relevant to the last research question of this paper "What steps have already been taken and what demands of actions are voiced?". "Modification from company", "rejecting" and "governmental actions" all have 21.05%, 15.79% and 10.53%, respectively (Fig. 8).

Comparing the metrics between the frequency in mention and the number of articles per tag, we found that their values are equal.

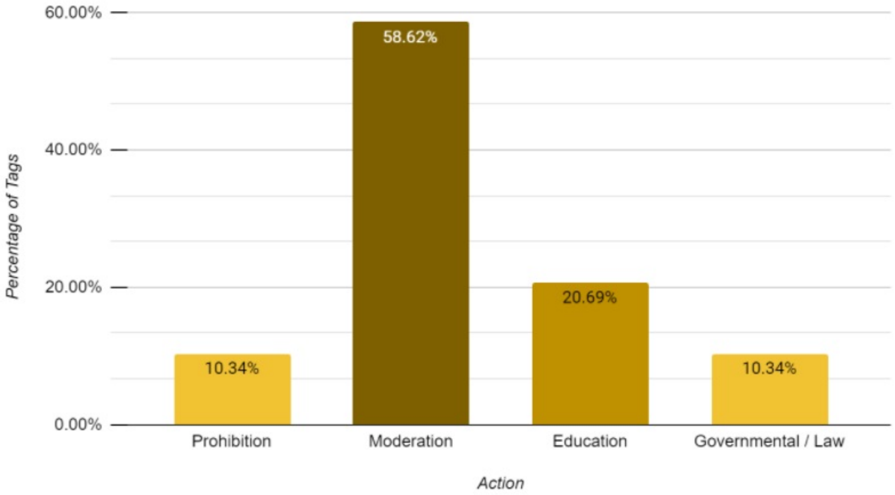


Figure 8: (Demand of) Actions for Generated Media

4 Discussion

4.1 Interpretation

Public discourses about the benefits, harms, effects, dangers of AI technologies are quite common among online magazines. For example, our results show that from 2018 to 2020 in the media certain specific public spheres were affected. These results indicate, first of all, that some areas of our life have such a strong influence on each other that any action in one area sets in motion another. This can be seen with politics and social media. Secondly, it should be noted that the influence of technology on a person has increased. It cannot be denied that a person is changing in the conditions of the existence of technologies that are aimed at distorting the truth, reality.

Also, our results were able to identify the most influential and discussed companies in recent years. The most influential and talked about companies in recent years on Facebook and Google com are found much more often than other companies, which means that they are more than anyone else responsible for the vector of development and perception of these technologies in public discourse. These important companies, along with the government, have a responsibility to work with each other to ensure the correct laws and procedures for the use of these technologies in social networks. We also think they need to educate people about fake news to prevent the unwanted resonance of this information. Consequences like this, usually lead to sad stories and ”rumors” in the articles that people are so eager to discuss.

The media regularly discuss ways to combat the impact of Artificial Intelligence technologies. People are mainly concerned with how to deal with an existing threat, delete information uploaded to the network, or change it. As already mentioned, they are not so worried about where this threat came from, our society forgets to ask what fake news is, and learn to perceive them correctly, as well as to deal with them before they arise. This also suggests that the media very rarely discusses how unique and useful the technology for creating synthetic media is, in contrast to personalization algorithms, which almost equally and positively and negatively affect today’s society.

4.2 Limitations

When dealing with news articles, it is always important to keep in mind that there are tendencies from the news outlets to report on negative topics. Even if positive articles are published, they statistically receive less readership and get shared less often on social media.(Leonhart 2021) The fact that we found a majority of articles to have a negative sentiment might therefore at least to a certain extent be traced back towards this statistical phenomenon.

IBM Watson appeared to be a far more problematic tool for analysing sentiment in newspaper articles than were expected.

Finding the right results with a keyword search can be difficult, since depending on the selected keywords, the results might differ a lot, even though one is still searching for the same thing.(Google 2009) That is why we tried to keep the search terms as neutral and as open as possible. However, the term “Deepfake” carries a negative connotation, as is demonstrated by *IBM Watson.* That connotation might influence the perceived sentiment against said technology.

There is some irony in the fact that we use a relevant feed as the tool for gathering our articles. Search engines like google do not show a standardized result and are inherently biased.(Google 2009) The question opens up on how proper scientific work is even possible with those tools since the results are not replicable due to the inconsistency of those algorithms. On the other hand it is worth questioning how far it is unavoidable or at least unrealistic to not rely on those algorithms for everyday life, considering the amount of data that each of us is being faced with every single day. A better question would be how and when to use them in a proper way or in a good way instead of whether we should use them or not.

5 Conclusion

A plethora of fake news and personalization algorithms are demolishing democracy, society, and public trust in government. Therefore, we have investigated how fake news and the dissemination of information spreads online, and also tried find a solution to detect these fake that had been invented by companies. The most influential people such as actors, queens, and political leaders have been affected by this fake propaganda. Most of the time, they use social media such as Facebook and Twitter. The study showed that political fake news spreads faster than other domains such as business and science. Hence, the companies are trying to find out the ultimate solution to cope with a common problem. One article which had been published on September, 2nd 2020 in Daily Mail stated that Microsoft has tried to disclose the videos and photos that have been manipulated in a bid to combat misinformation online. They have innovated a video authenticator tool that gives the confidence score to fake videos. The generative adversarial Network (GAN) algorithm is also being used to discriminate between real and fake news. The main goal was to answer the research questions defined in this study, which we partially coped with. Therefore, we have successfully used *Catma* to annotate, analyse, interpret and visualise.

References

Beard, Matt. 2019. *To fix the problem of deepfakes we must treat the cause, not the symptoms.* https://www.theguardian.com/commentisfree/2019/jul/23/to-fix-the-problem-of-deepfakes-we-must-treat-the-cause-not-the-symptoms (27 January, 2021).

Bhattarai, Abha. 2018. *The personal stylists who are training the bots to be personal stylists.* https://www.washingtonpost.com/business/economy/the-personal-stylists-who-are-training-the-bots-to-be-personal-stylists/2018/08/17/69bb476a-9f1d-11e8-93e3-24d1703d2a7a_story.html (27 January, 2021).

Brockell, Gillian. 2018. *Dear tech companies, i don't want to see pregnancy ads after my child was stillborn.* https://www.washingtonpost.com/lifestyle/2018/12/12/dear-tech-companies-i-dont-want-see-pregnancy-ads-after-my-child-was-stillborn/ (27 January, 2021).

Brown, Tom B., Benjamin Mann & Nick Ryder. 2020. Language models are few-shot learners. *NeurIPS*.

Chadwick, Jonathan. 2020. *Elvis back from the dead? artificial intelligence is used to create eerie 'deepfake' pop songs that sound like they are being sung by deceased stars.* https://www.dailymail.co.uk/sciencetech/article-8933235/AI-creates-deepfake-songs-sound-like-theyre-performed-deceased-pop-stars.html (27 January, 2021).

Chattopadhyay, Tuhina. 2020. *AI & The 2020 Elections.* https://www.mantra.ai/blogs/ai-in-2020-election/ (2 March, 2021).

Chivers, Tom. 2019. *What do we do about deepfake video?* https://www.theguardian.com/technology/2019/jun/23/what-do-we-do-about-deepfake-video-ai-facebook (27 January, 2021).

Cole, Samantha. 2017. *Ai-assisted fake porn is here and we're all fucked.* https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn (5 March, 2021).

Edwards, Scott & Steven Livingston. 2018. *Fake news is about to get a lot worse. that will make it easier to violate human rights — and get away with it.* https://www.washingtonpost.com/news/monkey-cage/wp/2018/04/03/fake-news-is-about-to-get-a-lot-worse-that-will-make-it-easier-to-violate-human-rights-and-get-away-with-it/ (27 January, 2021).

Goodfellow, Ian J., Jean Pouget-Abadie & Mehdi Mirza. 2014. Generative adversarial nets. *Universit´e de Montr´eal.*

Google. 2009. *Personalized Search for everyone.* https://googleblog.blogspot.com/2009/12/personalized-search-for-everyone.html (27 March, 2021).

Gupta, Pankaj. 2013. Wtf: the who to follow service at twitter. *WWW '13.*

Harwell, Drew. 2018. *White house shares doctored video to support punishment of journalist jim acosta.* https://www.washingtonpost.com/technology/2018/11/08/white-house-shares-doctored-video-support-punishment-journalist-jim-acosta/ (27 January, 2021).

Harwell, Drew. 2019. *A face-scanning algorithm increasingly decides whether you deserve the job.* https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/ (27 January, 2021).

Klosowski, Thorin. 2019. *Personalization has failed us.* https://www.nytimes.com/2019/11/05/opinion/personalization-privacy.html (27 January, 2021).

Leonhart, Bruce. 2021. *Bad news bias.* https://www.nytimes.com/2021/03/24/briefing/boulder-shooting-george-segal-astrazeneca.html (29 March, 2021).

Liberatore, Stacy. 2018. *Tinder for tv: ai swiping app can recommend shows based on what you like and what your friends are watching.* https://www.dailymail.co.uk/sciencetech/article-3538690/Tinder-TV-AI-swiping-app-recommend-shows-based-like-friends-watching.html (27 January, 2021).

McCabe, David & Davey Alba. 2020. *Facebook says it will ban 'deepfakes'*. https://www.nytimes.com/2020/01/07/technology/facebook-says-it-will-ban-deepfakes.html (27 January, 2021).

Media, 4International. 2019. *2019 NEWSPAPER WEB RANKINGS*. https://www.4imn.com/top200/ (10 February, 2021).

Metz, Cade. 2019. *Internet companies prepare to fight the 'deepfake' future*. https://www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html (27 January, 2021).

Morrison, Ryan. 2020. *Deepfake detection: microsoft unveils a tool that can tell if a video or photo has been doctored in a bid to combat disinformation online*. https://www.dailymail.co.uk/sciencetech/article-8688865/Microsoft-unveils-new-tools-identify-deepfake-videos.html (27 January, 2021).

Naughton, John. 2020a. *From viral conspiracies to exam fiascos, algorithms come with serious side effects*. https://www.theguardian.com/technology/2020/sep/06/from-viral-conspiracies-to-exam-fiascos-algorithms-come-with-serious-side-effects (27 January, 2021).

Naughton, John. 2020b. *How amazon puts misinformation at the top of your reading list*. https://www.theguardian.com/commentisfree/2020/aug/08/amazon-algorithm-curated-misinformation-books-data (27 January, 2021).

Parkin, Simon. 2019. *The rise of the deepfake and the threat to democracy*. https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy (27 January, 2021).

Pinkstone, Joe. 2018. *Snapchat's makeover arrives: app rolls out redesign that separates 'friends' and 'discover' pages to create a more personalised experience*. https://www.dailymail.co.uk/sciencetech/article-5361959/Snapchat-redesign-creates-personalised-experience.html (27 January, 2021).

Roose, Kevin. 2020. *Is tiktok a good buy? depends on what is included*. https://www.nytimes.com/2020/08/05/technology/tiktok-deal-algorithm.html (27 January, 2021).

Rosenberg, Matthew, Nicholas Confessore & Carole Cadwalladr. 2018. *How trump consultants exploited the facebook data of millions*. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html (5 March, 2021).

Schmidt, Andréa. 2019. *Deepfakes — believe at your own risk*. https://www.nytimes.com/2019/11/22/the-weekly/deepfake-joe-rogan.html (27 January, 2021).

Sharples, Eleanor. 2020. *Channel 4 will use a 'deepfake' version of the queen to deliver their version of her christmas message*. https://www.dailymail.co.uk/

news/article-9083401/Channel-4-mock-Queen-using-deepfake-version-Majesty.html (27 January, 2021).

Stefanou, Eleni. 2018. *Facebook's news feed change won't help social media addiction.* https://www.theguardian.com/commentisfree/2018/jan/15/facebook-news-feed-algorithm-social-media-addiction (27 January, 2021).

Vaccari, Cristian & Andrew Chadwick. 2020. *'deepfakes' are here. these deceptive videos erode trust in all news media.* https://www.washingtonpost.com/politics/2020/05/28/deepfakes-are-here-these-deceptive-videos-erode-trust-all-news-media/ (27 January, 2021).

Yang, Lin. 2018. Research on application of artificial intelligence based on big data background in computer network technolog. *OP Conf. Series: Materials Science and Engineering.*

Chapter 13

Does AI in public discourse change with different political and socio-economic systems? An analysis of the AI debate in newspapers in the emergent AI Superpowers: USA, China and Germany

Sabine Scholle, Konstantin Strömel, Archana Singh, Louisa Maubach, Johanna Kopetsch, Kyra Breidbach, Kristin Gnadt, Anna Ricarda Luther, Lea Tiyavorabun & Hedye Tayebi

China is one of the worlds AI Superpowers, yet we are prone to ignore the discourse the public is presented with in our Western-centric seminar. This study analyses popular newspapers of Germany, the USA and China and conducts a sentiment and SWOT analysis, to compare the style of discourse people of these nations are presented with. This study finds that China is the most positive in its discourse around AI, while the US and Germany mainly communicate information on artificial intelligence in an informative manner. In all three countries, the newspapers mostly cover favourable attributes of AI such as opportunities and strengths rather than adverse aspects like threats and weaknesses.

Keywords: artificial intelligence | newspaper | discourse | sentiment | SWOT | media | international | China | Germany | USA | comparative analysis

1 Introduction

The increase of practical applications and the ubiquity of artificial intelligence (AI) in our everyday lives, science and media (Fast & Horvitz 2017, Fischer et al.

2021) is undeniable. Media coverage, specifically news media coverage, plays an important role in the perception (Dominique Brossard 2013) and acceptance of AI of the general public. Members of the public are key factors for technology adoption on one hand due to their role as consumers and on the other hand as parts of the political system by the virtue of their citizenship. Thus, the general public has the power to affect the future of AI, which further highlights the importance to investigate how issues of AI are covered in the media.

Previous research examining the public perception of AI in the UK revealed that 25% of the participants considered the term AI to be synonymous to robots and that the most common visions of future impact of AI evoke fear and the sense of loss of control (Cave et al. 2019).These results emphasise the need for correct information to the public and "collaboration and inclusion of ethics and AI experts" (Ouchchy et al. 2020) in the public discourse. Moreover, since the survey was conducted in the UK, the results raise the question if, and if so to what extent, general perception and opinions as well as media coverage of AI varies between different countries and political systems.

Chuan and colleagues conducted a content analysis on a corpus of five major USA newspapers from 2009-2018 and found a predominant focus on business and technology. Furthermore, the benefits of AI were discussed more frequently, while the risks were discussed less frequently (Ching-Hua Chuan 2019). This is further substantiated by a recent study examining if media coverage of the topic AI and health care in the US displays a negative bias. The sentiment analysis is based on articles from 1958-2018 predominantly published in *The New York Times* and yields no evidence for a negative media coverage (Garvey & Maskal 2020). A similar pattern was obtained by an analysis of text corpora of *The New York Times* over 30 years of time, showing a consistently more optimistic discussion surrounding AI nowadays in the US (Fast & Horvitz 2017). A similar picture is painted by current studies concerning the public discourse on AI in Germany. Analysing the media coverage for the past 15 years, the study reveals the tendency of an over representation of economic and technological topics with the majority of media coverage being positive (Fischer et al. 2021). Furthermore, the media coverage in Germany increasingly lacks the discussion of the potential for the common good.

In contrast to the rather large body of research on the public discourse on AI in the USA and Germany, little research on the media coverage of AI in China was done. One current study compared the narrative on People's Daily Online with t he public discussion on the social platform WeChat, expecting a more critical debate on the social media platform (Zeng et al. 2020). Contrary to the hypothesis, the discourse on WeChat, like the discourse on People's Daily Online, is dominated by industrial and political actors and focuses on the economic potential while neglecting a critical debate (Zeng et al. 2020).

AI could have enormous consequences for our daily lives as individuals and as societies, therefore this can be seen as a question of how we see and will ultimately shape and live our future. Recognising the power that discourse yields by creating a framework within which people think about AI, we want to draw on and add to the developing body of research. The previous body of research touching on the media discourse on AI has largely analysed only one country for example by comparing multiple newspapers (Ching-Hua Chuan 2019) or studying one newspaper within a very long time frame (Garvey & Maskal 2020). Therefore, we aim to add a comparison of the media discourse surrounding AI spanning multiple world-leading countries (Xu 2019). From this more direct comparison we hope to see if there are tendencies in the respective discourse, which could be influenced by the economies, style of government or similar variables. In the scope of our work we only approximate this, by analysing one to three widely circulated newspapers in these three countries regarding AI research (Xu 2019) Germany, the USA and China. We aim to analyse comparable newspapers and chose the following: *The New York Times* (USA); *Zeit online*, *Süddeutsche Zeitung* and *Frankfurter Allgemeine Zeitung* (Germany); *China Daily* (China). Since one German source did not provide enough relevant articles, the German group included three distinct newspapers. Moreover, to have a comparable analysis, we tag articles selected by the same criteria and with the same ontologies. Due to our rather large number of articles, roughly 60 articles per country, we chose a quantitative approach with a combination of a sentiment analysis and SWOT analysis.

Due to Germany and Europe being a leader in regulatory practices affecting the technology sector (Meyer 2021, Kelion 2020), we expect to see in our analysis, that German newspapers report more critically on AI and highlighting potential threats and downsides, compared to the US and China. *China Daily* is a government-owned newspaper and the Chinese government recently devised a 150 billion dollar plan to become the world leading AI power, but keeps lacking effective regulation in the technology sector (Roberts et al. 2020). Therefore we expect *China Daily* to largely mirror the government's embrace of new technology and to emphasise its potential. The North American region, of which the USA is a substantial part, has 47% share of the AI market and is projected to grow (Analytics Insight 2020). Additionally, it's political and economic approach supports a relatively free market. Therefore, we expect *The New York Times*, even though it is a rather liberal newspaper, to be less critical than the selected German newspapers, but more critical than *China Daily* due to freedom of the press and independent journalism.

The results of our analysis can by no means be used to extrapolate to the whole media discourse surrounding AI in the respective country, but they are a first indicator and add some information to this very urgent research question.

2 Methodology

2.1 Data acquisition

Our group, consisting of ten members was divided into three sub-groups, one dedicated to the analysis of each country, Germany, China and the USA. Every sub-group analysed approximately 60 articles from popular newspapers in each country. The articles were selected based on their relevance and their publishing date being within the past two years. The German group analysed articles from *Zeit Online*, *Frankfurter Allgemeine Zeitung* and *Süddeutsche Zeitung*, while the Chinese group looked at articles from *China Daily* and the USA group covered articles from *The New York Times*. All newspapers are national newspapers of national relevance and both *The New York Times* as well as *China Daily* have international readership. All articles were taken from the newspapers' websites not from their printed form.

2.2 Procedure

The articles were tagged via CATMA, a computer assisted text markup and analysis tool. Two different tagsets were used in the markup process: SWOT and sentiment analysis. The former included the specified SWOT tags (*strength, weakness, opportunity, threat*) with the addition of *ethical concerns*, which was also covered with the tag *threat*. The tagging-style referred as close as possible to the agreed ontology to minimise subjective tagging behaviour. The sentiment analysis comprised three distinct tags: *positive*, *negative* and *neutral*, which were used to mark words or phrases that appeared in the context of an AI. The label *positive* was used to highlight words with a positive connotation and could therefore bias the reader towards a more positive view of AI while *negative* was used for words with a negative connotation that could create a negative bias towards AI. The tag *neutral* was made use of when the author provided unbiased explanations or information on AI systems and did not use a specific tone when mentioning AI. Each text had to offer a minimum amount of five tags to be taken into consideration for the analysis. All sub-groups used the same tagging procedure.

2.3 Data Analysis

The results were divided into two separate categories to perform two distinct analyses, one being the sentiment analysis and the other depicting the SWOT analysis. Then the proportion of each tag category per article inside the respective category was calculated in percentages. Since the reader usually looks at single articles and thus is only presented with the information that is provided in the given article, each article was assigned the same weight, despite variation in length. After obtaining the tag frequency per document, the mean frequency of each tag was computed for the interpretation of the general results. As articles often varied in topics and tone, it was essential to include a measure of dispersion, the standard deviation. For the comparison of the results of individual countries, a t-test was conducted for all possible country pairs. A result was considered significant with a p-value of less than 0.05.

3 Analysis of German newspapers

3.1 Thesis

In order to analyse the discourse about AI in Germany and be able to somewhat reflect the public stance of the German society towards AI, we base our analyses on three distinct German newspapers. Firstly, *Frankfurter Allgemeine Zeitung* which can be considered rather conservative and has a reach of roughly 0.83 million readers (Wikipedia 2021a). Secondly, two more liberal newspapers, *Süddeutsche Zeitung* with around 1.28 million (Wikipedia 2021b) readers and *Zeit online* with approximately 75.1 million website visits in January 2019 (Wikipedia 2021c).

As Germany is a democratic state with freedom of press and freedom of speech, we assumed that the articles would discuss the matter of artificial intelligence critically. Given the AI-strategy of the German Federal Government, the potential of AI should shape the areas of life and work in a safer, more efficient and in a more sustainable way (Bundesregierung 2021). Nevertheless, AI in Germany should be flanked by a strong level of IT-security (Bundesregierung 2021). Here the strategy focuses on the potential and opportunities AI offers, but still emphasises the importance of IT-security, considering the existing threats. Exactly this position is what we also expected to encounter during our analysis of the public discourse, represented by the news articles dealing with AI. As the German Federal Government intends to inform about the possible threats in its strategy, but still emphasises the areas in which AI offers potential of progress, we also expected to find a significant amount of neutral and informative parts in the articles.

To get a more detailed impression of Germany's AI strategy, Chapter 1 gives valuable insides to the government's paper *AI made in Germany: Strategie Künstliche Intelligenz der Bundesregierung*. The group conducted a quantitative and qualitative analysis investigating the way AI is communicated by the government and examining which of the main areas (general public, research, economy) the government dedicates the greatest attention to in respect of AI.

Regarding the three newspapers, which form the basis of our analysis, we expected the more conservative *Frankfurter Allgemeine Zeitung* to display slightly more sceptical tendencies. Whereas we anticipated that *Süddeutsche Zeitung* and *Zeit Online* as rather liberal newspapers lean slightly more towards the positive aspects of AI, namely opportunities and strengths.

3.2 Analysis of Results

We analysed 62 articles in total. Two articles needed to be excluded, since one of them had less tags than our minimum of five tags and the other one was mainly engaged in technology not artificial intelligence and therefore was off topic. Of these 60 articles we used 30 articles from *Frankfurter Allgemeine Zeitung*, 23 from *Zeit online* and seven from *Süddeutsche Zeitung*. The following table (Table 1) shows the means and standard deviations, first of the sentiment analysis and then of the SWOT analysis.

Table 1: Mean and StDev of all tags (Germany)

	Negative	Neutral	Positive	Opportunity	Strength	Threat	Weakness
Mean	0.253	0.448	0.2997	0.3347	0.2692	0.2047	0.192
StDev	0.2196	0.1938	0.1908	0.2291	0.1855	0.2011	0.201

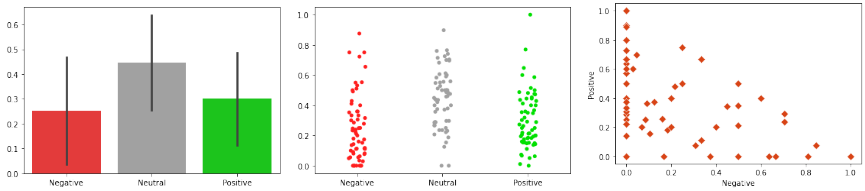
3.2.1 Sentiment Analysis

We found that the tone of the articles is predominantly neutral (44.7%) while biased statements, positive or negative, were more balanced with a slight tendency towards positive tone (*negative*: 25.3%, *positive*: 30%).

Figure 1a visualises the means and standard deviations. The height of the bars represents the mean values. The black lines illustrate the range of the standard deviation. It appears that *neutral* tags are most present. The line representing the standard deviation shows that the values of the articles with the lowest amount of *neutral* tags do not fall below a value above 0.2. Furthermore, the mean of the *positive* tags is slightly higher than for the *negative* tags and deviates from the mean in a range between 0.1 to 0.5.

Figure 1b illustrates the dispersion. Every dot represents one article and the value of its standard deviation. Observable is that the articles in the category *negative* stick the closest to the mean. In the category *positive* they differ more and in the category *neutral* the most.

Figure 1c displays the correlation of *negative* and *positive* tags. It portrays that many articles are solely positive in varying degrees whereas just a few are solely negative. Almost no article is evenly balanced but there are articles covering both, positive and negative sentiments.



(a) Tag frequency with StDev (b) Dispersion of tags (c) Correlation of neg/pos

Figure 1: Sentiment

3.2.2 SWOT Analysis

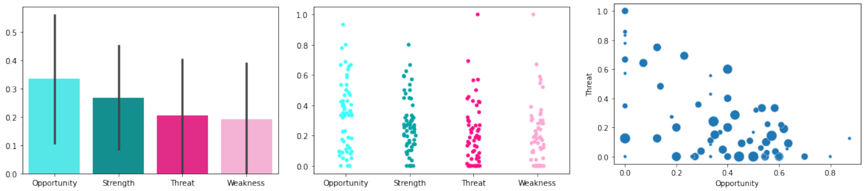
Within the SWOT analysis we find a higher proportion for positive aspects of AI represented by *opportunity* (33.5%) and *strength* (26.9%) as opposed to its negative aspects, *threat* (20.4%) and *weakness* (19.2%). The standard deviation is pretty high with a value around 20% across all tags, revealing that the proportion of tags highly varies across all gathered documents.

Figure 2a shows that the range of the standard deviation is high. Opportunities are most frequently discussed in the analysed articles since the mean frequency is highest for *opportunity*. The standard deviation however shows that the number of tags varies a lot across articles. *Strength* tags differ less according to the low standard deviation. Especially the categories *threat* and *weakness* include articles which have no *threat* or *weakness* tags at all and other articles which entail a greater number of these tags.

In Figure 2b the dispersion illustrates that the values for the tag *opportunity* differ the most from the mean with a rather equal distribution across the whole scale.

This displays that the articles are very different in the deviation to the mean. In the category *strength* the standard deviation is more concentrated around a value of 0.25. The same applies to the categories *threat* and *weakness*, with *weakness* being concentrated even more closely around its mean than *threat*.

In Figure 2c we can observe the level of neutrality of the discourse on opportunities and threats. The size of the data points represents the level of neutrality. It is evident that opportunities are more frequently communicated in a neutral manner than threats.



(a) Tag frequency with StDev

(b) Dispersion of tags

(c) Neutrality of Opp/Thr

Figure 2: SWOT

Finally, we also performed a comparison between the newspapers that we used. Therefore, we contrasted conservative and liberal newspapers on the basis of the mean and the standard deviation of our tags. Here a noticeable difference can be observed. The tone measured by the mean frequency of the sentiment tagset is slightly more negative for the liberal newspaper as the mean frequency of the label *negative* is up to approximately 28.1% while the conservative newspaper reaches around 22.5%. They show comparable results for the tag *neutral* (conservative: 44.09%, liberal: 45.3%) but again differ by about 6.8% in respect of the label *positive*, with a greater mean regarding the conservative source (conservative: 33.4%, liberal: 26.6%). The conservative newspaper also mentions opportunities more often than the liberal sources (conservative: 41.1%, liberal: 25.7%). However, in *Zeit Online* and *Süddeutsche Zeitung* strengths are more prevalent because the average frequency for these newspapers ranges up to 32% while for the *Frankfurter Allgemeine Zeitung* the mean frequency is only at 21.8%. Fewer threats can be detected on the conservative side (18.9%) than on the liberal side (22%) as well as a similar tendency regarding the category *weakness* (conservative: 18.2%, liberal: 20.2%).

3.2.3 Degree of Dispersion

The standard deviation is considerably high across all tags and both types of newspapers. Within the sentiment analysis of conservative newspapers values range from 19.3% for the *neutral* label to 22.1% for the *negative* tag. In the sentiment analysis of the liberal media values for the label *negative* are also most widely spread with a standard deviation of 21% while *positive* tags are most consistent and oscillate around 16.2%. Moreover, the results from the conservative newspaper in the SWOT category are the highest for *opportunity* (22.7%), the lowest for *strength* (13.7%). In contrast to that *opportunity* displays the lowest dispersion of roughly 19.7% in the liberal media whereas *threats* and *weaknesses* are comparably wide spread with a standard deviation reaching up to 21.5%.

3.3 Interpretation & Conclusion

Based on these results we can conclude that the analysed articles portray information about artificial intelligence to a great extent in a neutral way. They appear to set their focus on providing information accompanied with objective explanations of how artificial intelligence functions and how and where it is applied. The SWOT analysis also demonstrates that the analysed journalism also covers *opportunities* and *strengths* as well as *threats* and *weaknesses* of AI, proving that news coverage is not one-sided.

Yet, reporting is far from unbiased. Despite long sections of neutral information transmission there is usually still a message, which the author wants to convey about AI. Our results suggest that this message is overall more positive than negative as accentuations of positive aspects (*opportunity, strength*) can be observed in the German newspapers. It is interesting to see how the tone differs between conservative and liberal newspapers. Here we can generally perceive a more positive tone within the conservative source. Additionally, the liberal newspapers highlight slightly more negative aspects than the conservative one. Hence, as opposed to our thesis the conservative newspapers appear to be more open and less critical in their thematisation of artificial intelligence. The only outlier lies within the category *strength* where liberal newspapers have a higher share of tags than the conservative.

Concluding, the initial assumption claiming that German newspapers dedicate their main focus on providing information on AI in their reports have proven to be true, as the greatest majority of tags among the sentiment analysis is *neutral*. Additionally, the thesis that German newspaper present artificial intelligence in a critical manner can not be confirmed entirely. Since there is a slight tendency towards a positive tone, as well as greater focus on the potential of AI rather than on the given risks. The analysis shows that the discourse is not as balanced as expected. Also, our thesis concerning the discourse of AI inside the different kinds of newspapers does not hold. Contrary to our beliefs the liberal newspapers seem to be more sceptical and critical of AI than the conservative source.

4 Analysis of The New York Times

4.1 Introduction to Newspaper and Country

This part of the analysis concentrates on the media discourse on Artificial Intelligence in the US, more specifically on the representation of AI in *The New York Times (NYT)*. The USA as a country has roughly 330 million inhabitants, the highest GDP worldwide with $21.92 trillion (International Monetary Fund 2020), is a nuclear power and is considered by many as one of the most influential countries in the world (Schmidt 2021). The political approach is relatively hands off in terms of company regulation (United States Embassy 2021) and it is home of one of the most well known and influential tech clusters worldwide, the silicon valley. The USA is home to the Big Five: Apple, Google, Microsoft, Amazon and Facebook. Due to their size and international standing, these companies wield a lot of economic and political power and significance (Manjoo 2017). Additionally, North America is the region with largest AI market share worldwide, namely 47%, and expected to keep growing (Analytics Insight 2020).

The New York Times is the 3rd most widely circulated newspaper in the US, with 6.5 million subscribers as of the second quarter of 2020, its monthly readers amount to 150 million. It has a quite international outlook with 1,450 reporters reporting from over 150 countries (The New York Times Company 2021). It is known to have a liberal leaning, as it endorsed every Democratic candidate since 1960 (Brennan 2012). The audience is, relative to other news media outlets in the US, rather young, male, highly educated, in a higher income bracket and self identifies as relatively liberal and Democratic leaning (Pew Research Center 2012). Interestingly *The New York Times* is owned in 5th generation, since 1896, by the Ochs-Sulzberger family (Levitz 2016), which owns a majority of the open voting rights shares. Shares with non restrictive voting rights went public in 1969 (Lucey 2010).

4.2 Thesis

Taking the aforementioned facts into account with consideration that the USA does support free-market capitalism, we can infer that not only are progress and growth existential for the economy but narratives around those must be of high cultural significance. The IT sector specifically is of high economic importance as it is nowadays the main growing sector for progress and development partly through the advancements in AI. Therefore we think that economic progress narratives are also heavily fuelled by the development and improvements in the AI sector due to the possible opportunities it can represent for the global economy, (e.g. by enabling higher efficiency etc.) and also for the social domain (e.g. influence on everyday life through AI technology). Another aspect we take into consideration is the fact that *The NYT* is, as mentioned, widely known to be a liberal or liberal leaning newspaper. This would correlate with the stances of most tech entrepreneurs who commonly condone liberal policies (with the exception of regulatory policies) according to (David Broockman & Malhotra 2017). Because of ideological alignments there might also be more alignments in the interpretation of events in the AI sector. Contemplating all of our thoughts, we conclude that we would expect *The NYT* reports on AI to be showcasing a higher tendency towards the representation of opportunities and strengths as opposed to threats and weaknesses as the positive aspects are thought to be met with more keenness and developments in general to be interpreted rather positively. Nevertheless, we expect the articles to be predominantly of neutral tone due to *The NYT* being an informative newspaper in which objectivity in the communication of events is anticipated.

4.3 Analysis of Results

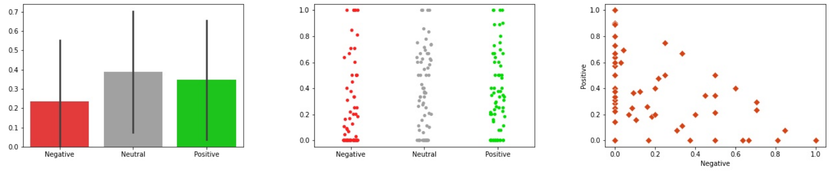
We read and tagged a total of 79 articles by *The New York Times*, of which 10 had fewer than 5 tags and therefore were not considered in our analysis. Left are 69 articles for the final analysis. The raw tag counts were transformed into percentages (of tags per article), where sentiment and SWOT tags are represented separately.

Table 2: Mean and StDev of all tags (USA)

	Negative	Neutral	Positive	Opportunity	Strength	Threat	Weakness
Mean	0.2368	0.3882	0.346	0.3761	0.2464	0.2287	0.1488
StDev	0.3189	0.3161	0.3102	0.2239	0.211	0.2748	0.175

4.3.1 Sentiment Analysis

The majority of sentiment tags are *neutral*, on average 39% per article. Followed by *positive* tags, making up roughly 35% of the tags, while on average only 24% of the tags in an article are *negative*. When adding the sentiment tags up, there are 2% missing. This difference occurs, because of the 69 articles analysed, two have no tag belonging to the sentiment tagset. The standard deviation lies between 0.31 and 0.32 for all three tags (Figure 3a). This rather high standard deviation as well as Figure 3b show that the tag percentages are highly dispersed among the articles. Additionally, in Figure 3c it can be seen that most *positive* tags accumulate where no *negative* tags occur, showcasing that not many articles have a balance between *negative* and *positive* tags.



(a) Tag frequency with StDev (b) Dispersion of tags (c) Correlation of neg/pos

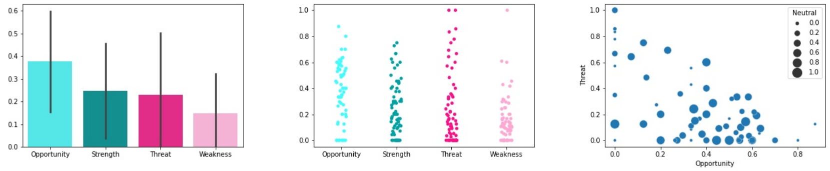
Figure 3: Sentiment

4.3.2 SWOT Analysis

The SWOT tag which occurs most often is *opportunity* with a mean of 38%. Followed by *strength* (25%) and *threat* (23%), which are used at a similar rate. *Weakness* is the rarest tag, only making up around 15% of all tags. The standard deviation is highest for *threat* tags (0.27). *Strength* and *opportunity* have a standard deviation of 0.21/0.22, *weakness* the lowest one with 0.17 (Figure 4a). We visualised this dispersion and how it differs amongst the SWOT tags in Figure 4b. There is an interesting correlation between the occurrence of *neutral* tags together with *opportunity/threat* tags in articles: Neutral tone occurs more often in articles underlining the opportunities of AI, rather than the threats posed by AI (Figure 4c).

4.3.3 General Observations

In many articles, the workings of the algorithms underlying AI are explained rather scientifically and in-depth considering *The New York Times* does not have an audience from a scientific background. These sections are reflected in the *neutral* tags. We also notice that most articles either address positive or negative aspects of AI and its use, which is supported by the high dispersion of tags as well as the correlation between *positive* and *negative* tags (Figure 3c). Thus, there are



(a) Tag frequency with StDev (b) Dispersion of tags (c) Neutrality of Opp/Thr

Figure 4: SWOT

barely any articles showing both sides of the argument with a balanced account of the prospects and risks of AI.

The context in which *threats* are touched upon in the articles, often concerns either loss of jobs, data security or racial (and other discriminatory) biases, or some general dystopian notion of how AI will affect our lives. Opportunities are mentioned mostly in combination with improvements regarding technology and the health sector (especially related to the fight against Covid-19), but also with respect to the creation of jobs and social opportunities.

4.4 Interpretation & Conclusion

The present data does support our hypothesis of a predominantly neutral tone in the discourse surrounding AI in the *The New York Times*. Even though, when comparing the amount of *positive* tags with the amount of *negative* tags a tendency towards a more positive media coverage is obtained, the vast majority of articles displays a primarily neutral reporting (Figure 2). Interestingly, all three tags exhibit a high dispersion (Figure 3b) hinting towards a tendency towards one tone category within the individual articles. This is further substantiated by Figure 3c, clearly visualising the accumulation of data points along the axes representing positive and negative tone. Thus, the data at hand yields no evidence of a negative bias (replicating previous results (Garvey & Maskal 2020)), or a positive bias of the media coverage of AI in the US while simultaneously pointing towards an tendentious tone within the individual articles.

By contrast, the SWOT-Analysis only partly supports our hypothesis of discourse mainly focusing on the opportunities of AI. The *opportunity* tag is predominantly used and makes up 37.61% of all SWOT tags, followed by the *strength* tag (Table 2). Hence, both tags concerned with the positive aspects of AI are used most often. This finding aligns with the liberal orientation of *The New York Times* and the silicon valley which could account for the similar interpretation of developments in the sector as an opportunity. Previous findings of a focus on economic and technological opportunities (Ching-Hua Chuan 2019) can be replicated, yet also the health sector and social issues are addressed. The threats are discussed roughly as often as the strengths and the least used tag with 17% is the *weakness* tag (Table 2), indicating the under-representation of negative aspects in the public discourse. Crucially, the *threat* tag exhibits the highest dispersion of all tags (Figure 4c), further hinting towards a strong tendency to primarily discuss one aspect of AI development within the individual articles. Furthermore, when comparing the *opportunity* and the *threat* tag, it is evident that articles mainly covering opportunities of AI are tending to be more neutral (Figure 4c). Even though the standard deviation for all tags is lower than within the tone analysis (Figure 3), they are still comparably high, further supporting the hypothesis of rather unbalanced reporting within the individual articles.

In summary, the present data supports our expectation of a neutral tone with a focus on the opportunities of AI while simultaneously pointing towards a lack of balanced media coverage in both sentiment and SWOT analysis.

5 Analysis of China Daily

5.1 Introduction to Newspaper and Country

China has rapidly transformed in recent years to be one of the world leaders in technological innovation and output. With a population of almost 1.4 billion people, not only is China the world's second largest economy, it also has the highest global internet penetration rate (over 800 million internet users). (Lam et al. 2019) Furthermore, being a communist country with long-ruling political leaders, China also has the advantage of being able to set and follow through on long-term strategies. A prime example of this is their 2017 state council plan aiming to become the world leader in AI by 2030. (Westerheide 2020) The government has invested heavily in AI with half of the world's total investments in AI coming from China. Furthermore China's triumvirate of tech giants 'BAT' (Baidu, Alibaba and Tencent) are also pushing the development of areas previously dominated by US-American companies.

China Daily is the biggest English-language newspaper in China with a circulation of 900,000 copies. Two thirds of these are distributed overseas (China Daily 2021b). Its readership consists mainly of foreigners and nationals working in high-end positions, namely diplomats and governmental policy makers (China Daily 2021a). This media outlet is owned by the Chinese Communist Party. *China Daily* provides information on chinese politics, economy, society and culture and they are often referred to as the "Voice of China" or "Window to China." (China

Daily 2021a) This is of significance as China ranks as 177th of 180 countries in terms of free press, having neither press freedom nor open access to the internet as well as exercising online surveillance of its citizens (World Press Freedom Index 2020). This allows President Xi Jinping to build a very curated public discourse on AI.

5.2 Thesis

Due to China's government having the aforementioned iron grip on information channels as well as its ambitious plans for AI dominance - it seems reasonable to infer that Chinese Communist Party would want to report on AI in a complimentary light. This would encourage the population to support AI innovation and investment - aligning with the goals of the party. Thus, the AI sentiment of *China Daily* articles is expected to be primarily *positive* (with modest *neutral* and minimal *negative* tags). Likewise in SWOT analysis - a heavy presence of the *strengths* and *opportunities* of AI is expected, avoiding *weaknesses* or *threats*.

5.3 Analysis of Results

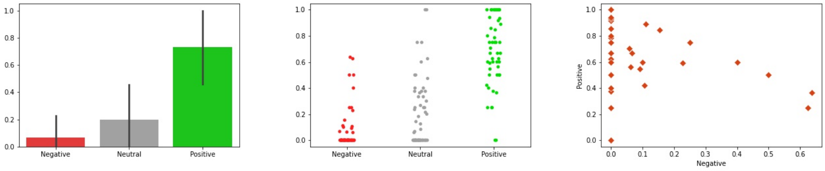
In total, out of 68 articles, two were removed because they were not specifically related to AI, but rather to 5G technology and another six were removed for having less than five tags. The remaining 60 articles were then analysed using the counts of each sentiment/SWOT tag.

Table 3: Mean and StDev of all tags (China)

	Negative	Neutral	Positive	Opportunity	Strength	Threat	Weakness
Mean	0.0701	0.1996	0.7302	0.5152	0.3273	0.0616	0.0959
StDev	0.1564	0.2599	0.271	0.2857	0.3027	0.1199	0.1741

5.3.1 Sentiment Analysis

Analysis of the sentiment results show a remarkable 73% *positive* tags, 19% *neutral* and a mere 7% that are *negative* (Figure 5a). The standard deviation lays between 0.15 *(negative)* to 0.27 *(positive)*.This is interesting as it matches the order of the mean number of tags - showing that there is a direct positive correlation - and could be because the more often a sentiment-type is tagged, the more liable it is to deviation. The dispersion of tags in Figure 5b, indicates an unbalanced tendency of positive articles to be hugely so, whilst the opposite is true with negative articles which are grouped together at the lower end of the spectrum, and finally neutral articles are seen to be slightly more balanced, but also grouped more towards the y-axis. The indirect *positive/negative* correlation of Figure 5c, with the high dispersion of points clustered along the y-axis shows that highly positive articles are generally not negative in the slightest.



(a) Tag frequency with StDev

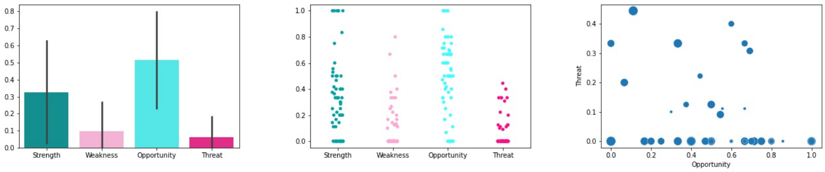
(b) Dispersion of tags

(c) Correlation of neg/pos

Figure 5: Sentiment

5.3.2 SWOT Analysis

In the SWOT analysis, it is clear that *opportunity* is tagged the most with over half (51%). *Strength* follows with 33%, *weakness* next with 9.5% and finally *threat* with only 6%. Standard deviation lays between 0.12 *(threat)* and 0.3 *(strength)*. Again, as seen in the sentiment analysis, there is a correlation between number of tags and the standard deviation - supporting the concept that the more predominant a SWOT tag is seen in articles, the more deviation it exhibits (Figure 6a). Figure 6b also shows a strong dispersion of *strength* and *opportunity*, however not for *weakness* or *threat* that are clustered further towards the y-axis. There is also a noteworthy lack of neutrality between *opportunity* and *threat*, most articles being singularly *opportunity*-biased.



(a) Tag frequency with StDev

(b) Dispersion of tags

(c) Neutrality of Opp/Thr

Figure 6: SWOT

5.3.3 General Observations

An extremely clear positive stance on the current and planned dominance of AI in China, as well as it's current and future benefits, are obvious from the onset. To allow the reader a first-hand understanding of this, below are a few article excerpts conveying the tone and type of reporting. For example, a 2019 article titled 'Artificial intelligence adoption gathering momentum in China' is quoted reporting, "Over the last decade, China has made remarkable progress to become the world's artificial intelligence powerhouse. Advances in data collection and algorithms, and the prevalence of mobile devices, coupled with aggressive R&D spending, have helped the country achieve major breakthroughs in the AI realm," (Lee 2019). This type of almost utopian-esque portrayal is prevalent throughout *China Daily's* journalism.

Another quote from the same article shows the polarisation of ethical standpoints across the countries on a topic like privacy concerns, "Chinese business leaders are also hoping for regulatory relaxation. More than 80 percent of the respondents agreed that the government should limit data collection regulations to facilitate AI development."(Lee 2019). Whereas most Western countries are seen to consider privacy concerns an AI threat, China's discourse seems to be not only less concerned, but actually advocating the lessening of regulations.

Lastly, in the unusual case when an article may touch on something negative, such as in the opinion piece "Artificial intelligence boom continues despite mistrust" by David Lee, then *China Daily* generally negates its association with the author by mentioning, "Editor's note: David Lee is a consultant and [...] the article reflects the author's opinions, and not necessarily the views of CGTN." This is despite the fact that usually, the article is still remarkably pro-AI, as seen in this section from the same aforementioned article: "First of all, is AI taking away jobs? Yes, robotics backed by AI is starting to replace human labour in the manufacturing sector. However [...] if AI is replacing simple labour, why can't innovative societies just create more jobs that AI cannot replace? [...] Instead of blaming AI for "stealing" our jobs, why not blame ourselves for failing to innovate fast enough?" (Lee 2019).

These quotes are used to show the radical disconnect in tone of reporting between the three countries addressed in this paper. In most articles in *China Daily*, reporting is done with a single-sided view to promote the development of AI (supported by the dispersion graphs seen in Figure 5b) and Figure 6b)). Figure5c shows this clearly as well, in that almost no articles compared pros and cons of AI to deliver an all-round balanced piece.

More generally, the most often cited positive points address healthcare, education and better efficiency in *opportunity*-style tags and the improvement of the economy by helping businesses in *strength*-tags. In both *threat* and *weakness* - so few tags are present, that finding over-riding themes are difficult.

5.4 Interpretation & Conclusion

Analysis of the results points to support our hypothesis of a predominantly positive tone (Figure 5a) along with a strong focus on the *strengths* and *opportunities* of AI (Figure 6a) rather than the alternative, namely a focus on the *weaknesses* or *threats*. This is expected, as to align the public discourse of AI in China to support the views and ambitions of the ruling Chinese Communist Party and their leader Xi Jinping. The lack of neutral reporting, with almost no articles weighing up both sides of the argument, to allow the reader to make an unbiased, autonomous decision, is even higher than expected, as seen in the indirect correlation in Figure 5c, but not surprising looking at the tag frequency graph in Figure 5a.

In summation, we find strong support for our hypothesis and would like to reiterate that the data is dramatically skewed towards positive AI articles discussing *opportunities* and *strengths* which directly aligns with the Chinese Communist Party's goals. As the *China Daily* is a direct mouthpiece of the party, it can be safely inferred that the government is using this platform to further their ideals of becoming an AI superpower within the next decade.

6 Combined Results

To enable a clear comparison of our results, we applied a t-test between each pair of countries at a time. We tested for significant results (a p-value of less than 0.05) in both the sentiment and SWOT analysis. Moreover, we state the means which lead to either significance or insignificance.

6.1 Germany - USA

6.1.1 Sentiment Analysis

While comparing Germany versus USA we conduct a t-test and find no significant differences between those countries in the sentiment categories *negative* (mean: USA = 0.24, Germany = 0.25),*neutral* (mean: USA = 0.40, Germany = 0.45) and *positive* (mean: USA = 0.36, Germany = 0.3).

6.1.2 SWOT Analysis

In respect to the SWOT analysis our t-test results also show no significant results at all: *opportunity* (mean: USA = 0.38, Germany = 0.33), *threat* (mean: USA = 0.23, Germany = 0.2), *strength* (mean: USA = 0.25, Germany = 0.27) and *weakness* (mean: USA = 0.15, Germany = 0.19). By observing the means you can see that the USA obtain higher values in the categories *opportunity* and *threat* and Germany in the categories *strength* and *weakness*. Subsequently, the examination of positive and negative aspects is very balanced between this pair.

6.2 Germany - China

6.2.1 Sentiment Analysis

The comparison of China with Germany using a t-test reveals significant difference in all sentiment categories *negative* (mean: China = 0.70, Germany = 0.25), *neutral* (mean: China = 0.2, Germany = 0.45) and *positive* (mean: China = 0.73, Germany = 0.3).

6.2.2 SWOT Analysis

The SWOT categories differ in the majority of categories significantly: *opportunity* (mean: China = 0.52, Germany = 0.33), *weakness* (mean: China = 0.1, Germany = 0.19) and *threat* (mean: China = 0.06, Germany = 0.2). The only category which yields comparable results between Germany and China was the category *strength* (mean: China = 0.33, Germany = 0.27).

6.3 USA - China

6.3.1 Sentiment Analysis

Lastly, we contrast China and the USA. Since Germany and the USA vary not much, the results of USA versus China are very similar to the results comparing Germany and China. The sentiment analysis leads to significant findings in all categories: *negative* (mean: China = 0.07, USA = 0.24), *neutral* (mean: China = 0.2, USA = 0.4) and *positive* (mean: China = 0.73, USA = 0.36).

6.3.2 SWOT Analysis

The same applies to the categories *threat* (mean: China = 0.06, USA = 0.23) and *opportunity* (mean: China = 0.52, USA = 0.38). In this comparison, unlike to Germany versus China, not only *strength* (mean: China = 0.33, USA = 0.25) but also *weakness* (mean: China = 0.1, USA = 0.15) exhibits insignificant t-test results.

6.4 USA - China - Germany

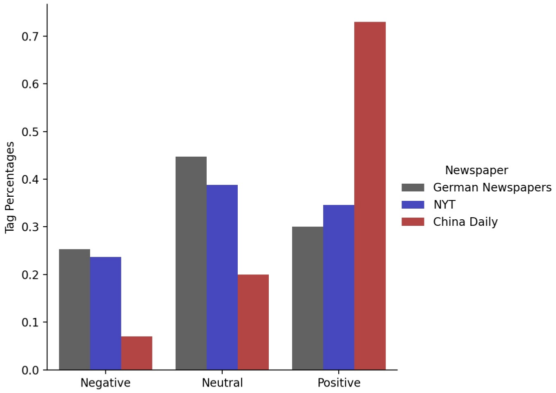


Figure 7: Means of Germany - USA - China

6.4.1 Sentiment Analysis

Overall *China Daily* has the most positive attitude towards AI. The newspaper uses positive descriptions approximately 73% more than USA (36%) and Germany (30%) , which do not report in such a positive manner. Even though the t-test reveals no significance it might be notable that the USA tends to report more positive than Germany. Moreover, *China Daily* avoids reporting in a negative manner and does so far less than the German and US-American newspapers (China 7%, Germany 25%, USA 24%).

Accordingly, the neutral media coverage is the least within *China Daily* 20%, whereas the USA and Germany exhibit nearly the same values, the former 40% and the latter 45%.

6.4.2 SWOT Analysis

Similarly, positive ascriptions as *opportunity* and *strength* are emphasised the most often within *China Daily* (*opportunity* = 52%, *strength* = 33%). The category *opportunity* varies considerable from Germany (*opportunity* = 33%) and the USA (*opportunity* = 38%). But the differences in terms of the category *strength* are not big enough to be mentioned as significant (Germany = 25%, USA = 27%).

Weaknesses are stated approximately equally often between USA and China, but vary significantly between Germany (19%) and China (10%). And after all, the results within the category *threat* are about 20% for USA and Germany, but for China just 6%.

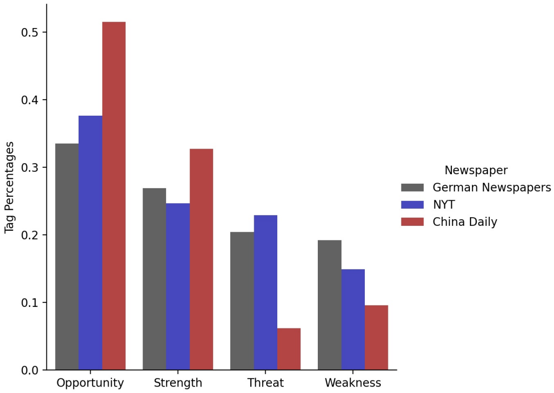


Figure 8: Means of Germany - USA - China

7 Discussion

7.1 Interpretation

This section addresses whether the results align with our more general expectations on the comparison between-countries sentiment and SWOT. In the paper's introductory thesis we mentioned that we anticipate seeing German newspapers reporting more critically on AI, highlighting the threats thereof and its downsides in comparison to the Chinese and USA newspapers. Whereas in order for *China Daily* to mirror the Chinese government's embrace of technology, it was expected to emphasise its potential and strengths. Finally, with the USA being a free market economy with the largest stake in the AI market, *The New York Times* was expected to report more positively than Germany, however still more critically or more balanced than China due to more possible leniency and mitigated liberties in the press.

The data suggests that these expectations were completely in line with the sentiment results but slightly different in the SWOT results. China definitely displayed the most positive attitude towards AI, with USA following and Germany in last place. This order is reversed with the *negative* analysis, Germany being the most uncomplimentary, then the USA (although worth mentioning only by 1%) and finally China far behind.

In terms of SWOT analysis, the AI descriptions of *strengths* and *opportunities* were again most predominant in China, with the USA describing *opportunities* more often than Germany, but both the USA and Germany referring to *strengths* in the same number. This is a slight deviation from our expectations since the USA was predicted to be more open to discussing and presenting AI's strengths rather than Germany. Then, China and the USA reported on *weaknesses* in almost the same fashion, whereas Germany does so in a much higher amount. Finally, *threats* are addressed almost equally between the USA and Germany but again far less in China. Again here, it was expected that Germany would have a higher count of *threats* than the other countries but instead our analysis revealed a similar amount in presentation of *threats* as the USA reporting.

Even with these deviations, the data provides convincing substantiation that differing political and economic systems affect public discourse on AI. This is a significant issue in a world divided by country, political and cultural borders, yet relying on AI and international co-operation to usher us into a new phase of development and modernism. We need to work together in a global fashion as never before, yet clearly have vastly differing sentiments and views on the very technology that we are using to do it with.

The race for AI dominance and a global innovation advantage is fuelled by the underlying economic interests. There is no unilateral approach to which strategy will lead to winning the race. But surely the factors of the financial funding and support of new tech, as well as raising skilled personal or gathering the most valuable data, will play their role in the development of the most dominant country regarding AI. China has great preconditions to become the world leading nation to use AI for their economic progress (Westerheide 2020). The large population with almost one billion internet users produces a huge amount of data which can be effectively used (Cheng 2021). There might be an interest from the government's side to establish a predominantly positive discourse towards AI, in order to prevent a sceptic society from slowing down the economic progress which is achieved through the application and development of AI. As *China Daily* is owned by the Chinese Communist party, the approach to eliminate as many negative tendencies as possible from the public discourse is reflected in the analysis results.

AI is also one of the top priorities on the political agenda of the USA, as it is home to many of the largest, most influential tech companies. The economic interests in the further development of AI is apparent. Nevertheless, the discourse in *The New York Times* is not directly influenced by the federal government, opposed to *China Daily*, since it is a privately owned newspaper. Furthermore, the USA ranks on place 45, unlike China on place 173 out of 180 countries, concerning the freedom of press (World Press Freedom Index 2020). This fact as well as the USA having a federal government paves the way for an critical discourse regarding AI (The White House 2021). Besides that, *The New York Times* as a rather liberal newspaper with a quite international outlook might support an open discourse that highlights different aspects of AI, which is reflected in the analysis results.

As Germany ranks on place 11 out of 180 countries in respect to their freedom of press it does not come as a surprise that there is an open discourse around AI highlighting different aspects from the entire spectrum (World Press Freedom Index 2020). Nevertheless, economic interests also fuel the further development in the sector of these technologies in Germany. But the debate surrounding AI also evokes serious concerns on different levels, one being data security. The German government takes up on these concerns by emphasising in the goals of their AI strategy, that ”data in Germany should solely be used for the benefit of society, environment, economy and the state” (Bundesregierung 2021). Further, it is also stated in the strategy goals, that ”Artificial Intelligence made in Germany aims to become a world wide known seal of excellence.” (Bundesregierung 2021), which suggests that the basis of success will also be the trust of society in the application of these new technologies.

Otherwise, other observations we made include that *The New York Times* and *China Daily* tend to do event-based reporting whereas the German newspapers produce more opinion-type pieces and updates from the tech industry as a general discussion and contemplation of the topic without a specific occasion or incident inspiring an article. A possible conclusion can be derived from our thesis that German AI-discourse can gravitate towards fore-thoughtfulness. Then, before moving on to further interpreting the newspapers’ tones and content, we ought to acknowledge that *The New York Times* in comparison to the German newspapers and *China Daily* has more sensational-style, person-driven stories surrounding the topic. A possible explanation is that a more captivating narrative is especially important for subscription-based newspapers to establish their finances as well as, could in accordance with the taste of American news-style reporting.This observation may encourage further interest and research in the cultural norms of news-telling in the societies, but evidently this ties into deeper inspections into culture and mass psychology.

What also shapes public discourse and is indicative of prevalent cultural narratives would be the used tonality in the reporting and the selection of topics and themes regarding AI. The central question here is ”What is presented and how is it presented?”

Contemplating our already discussed findings, we summarise that the tone in the German newspapers and *The New York Times* was largely neutral while it was predominantly positive in *China Daily*. In the interpretation of the individual results it was concluded that the positive tone in *China Daily* aligns with the long-term economic plan of rapidly advancing AI technology of the Chinese Communist Party. In contrast, in German and US-American newspapers it is less clear whose interest is behind the reporting or even what interest, if there is one other than the simple retelling of events, shapes the media. As for the provision of information, objectivity (as far as possible) in news report is socially expected and the norms in journalism have adjusted accordingly. These findings and interpretations open up further discussion regarding the public's relation to the media (and the involved socio-economic and socio-political interplay) which would be beyond the scope of our analysis.

Moving on from the ”How?” to the ”What?” we also pay attention to the specific content of our tags, i.e. what is being described here as an *opportunity* / *strength* or *threat* / *weakness*? The interpretation is evidently influenced by the tone as technically an observation can be described and consequently be read as an opportunity if the tone and the wording is positive and vice versa. However, we argue that it is possible to keep a neutral tone while outlining something that would be viewed as positive/negative not in virtue of chosen phrasing or reporting style but in virtue of cultural values and common socially adjusted thinking patterns that have historically emerged for a society and are true for one's living reality. To further clarify, consider that just writing out the statement: ”The development of Artificial Intelligence will cost some jobs in the future.” is not worded negatively but its content is arguably regarded as a possible threat of AI since our material well-being is tied to our employment and it is also common practice to measure economic positions on unemployment rates for instance (again, this would indicate that the international economic position of a country is important for the population since it is also arguably tied to a country's material conditions).

We suggest that the content of the presented *strengths*, *weaknesses*, *opportunities* and *threats* are shaped by and do shape public discourse surrounding AI while simultaneously showcasing what socio-cultural and political issues are primarily considered in the discussion of the topic. It is of interest to recognise what the primary concerns of a country are when AI is evolving and what it is revealed through that analysis. Again, the extent of those factors are debatable but nevertheless, we are able to interpret our aforementioned findings. For example, *strength* and *weakness* for the most part are either concerned with the rapid development of AI as a *strength* or its shortcomings (errors, storage issues etc.) as a *weakness. The New York Times*' reports vary more greatly in *opportunity* as opposed to *threat*s which primarily focus on algorithmic biases. The addressed issues are particularly racial biases although throughout the articles multiple societal groups are said to be disadvantaged or ignored by AI whether it be gender biases, biases against people with disabilities etc. While these threats are acknowledged by all newspapers, the uniformity of the content of what we tag to be a *threat* in *The New York Times* peculiarly dealing with this notion is an interesting finding. It elucidates current US-American social discussions and highlights the difficulties the society has regarding racism. Observations like these are to be expected taking into account our impression of the development of public discourse contemplating and discussions ideas surrounding social justice which we see being done globally but primarily stemming from the USA. With this mind, we could conclude that this reveals that a more thorough knowledge on statistics and data gathering as well as discussion in the AI sector is needed to avoid biases or to be even detected. Furthermore, this leaves open the question of the necessity of enhancement of working places creating AI.

The previously mentioned *threat* regarding the loss of jobs also comes up in every newspaper. Although *The New York Times* and *China Daily* often mitigate this threat by communicating that new jobs will also be created, we would argue that this commonly felt threat is indicative of structural issues where economic progress that is systematically needed for a population's well-being paradoxically puts people into a disadvantage as an inevitable effect. We do suggest that a positive presentation of AI development fits economic progress narratives but this presented threat equally is line with the put importance on economic well-being as it directly deals with the topic. The next commonly considered *threat*, especially by the German newspapers, is the opacity or otherwise poorly communicated strategy regarding data security. Because of that it might be possible that more transparent and thorough discussion about regulations will inevitably be needed. *China Daily* addresses this issue in a lower frequency which we interpret to be a result of a more centralised method of regulatory controlling mechanisms while in contrast, the data collection being structurally and more commonly privatised in the West, complicates the issue in a different way. The last universally felt *threat* that we observe is the fear of humanity not catching up with AI and the machines taking over. It ties in with the fear of the loss of jobs but arguably also reveals a deeper fear of losing control and a sceptical view on technology. We observe this primarily in the German newspapers which would be in line with our thesis of Germany being more cautious and conservative in its approach to technological development and its applications.

To address a universally described *opportunity*, we additionally recognise how AI is presented as an *opportunity* to combat loneliness. Setting our doubts on the fulfilling quality of socialising with an AI aside, we also find it to be symptomatic for a society's problem with loneliness and isolation. In addition, it is indicative of a society's normalisation of evolving more extensive commodification of aspects of life like that of relationships in late-stage capitalism. While the German newspapers also mitigate this framing by explicitly stating that AI does not replace human beings in that regard, *China Daily* does not deal with the topic as extensively. As for the German newspapers, it yet again conforms to our general conclusion the newspaper tends to be more vigilant in its reporting. Regarding China, a possible explanation we assume is that a potentially more collectivist identity shapes notions around those topics as well as AI being essentially tied to economic opportunities to render this specific *opportunity* to be less of a thought.

All in all, these considerations can further encourage the analysis and interpretations of our findings in regards to the cyclical relationship between the media and public discourse. However, the representative quality of our finding is rudimentary. This issue alongside further limitations of our study is discussed in the following section.

7.2 Limitations

Since we manually tagged the articles, the approach on how to apply the agreed ontology, might have differed among the group members. Therefore a possible individual bias in our tagging can not be excluded. In general, our personal sociocultural background and experience can have an influence on how we read and interpret the communicated ideas in the newspapers. This ties into the aforementioned notions of neutrally presented facts being viewed as *positive* or *negative* considering our living realities. The chosen newspapers are based in different countries with different backgrounds and therefore caution in the interpretation and subsequent tagging was applied. Furthermore, it is questionable if the chosen newspapers are sufficient to represent the public discourse in the respective countries. Firstly, *The New York Times* addresses a quite international audience and is arguably not a US-based newspaper the same way the German newspapers are for Germany, for instance. This specific instance was mitigated by choosing only US-American authors and primarily US-based reports. Then, *China Daily* is written in English principally for a non-Chinese audience. It therefore consequently excludes the non-English speaking Chinese majority (Song 2021) from having any access to the given information. It is, however, still a state-owned national paper with Chinese Communist Party political alignment. However, it is conspicuously nearly impossible to find an accurate representation for a whole country through the discourse on a newspaper since not only can the outlook within the articles themselves be divided but in every presented country, social, political and cultural differences are present making it difficult to summarise a country's overall attitude. Nevertheless, we suggest our analysis as a starting point for contemplation and find within the limitations (that would be present in any field) the chosen sources suffice for the purpose of comparison, since all of them are very influential and have a high significance in their country.

Additionally the proportion of each tag is calculated within each article. Therefore, equal weight is put on all articles regardless of their size and number of tags. This can pose a problem, since reading ten pages explaining in-depth why artificial intelligence should be treated with caution, might be more convincing than reading two pages with the same intention. While it might be debatable that this treatment would yield the most accurate representation, due to the fact that nevertheless a starting point is needed and the selected newspaper suffice in their similarities and significance for comparison, we did still tag the different longer and shorter articles whose lengths tend to rely on the newspaper. Our reasoning is that by virtue of the existence of an article, it is representative of the shape of public discourse. We would argue that the amount of articles on AI showcases a high interest with many authors and occasional triggers being included in the creation of varying reports. Many shorter articles will be more prevalent with their points per article than one very long article with many points. Accordingly, we place a higher value on the proportion of tags within each article as opposed to the sum of all tags throughout all of the articles.

Tagging by hand gives us the chance to go in-depth with our analysis and apply our own individual interpretations on themes that a machine might not detect. However, that kind of manual approach also means that our capacity regarding the amount of tagged articles is smaller. Accordingly, it has to be noted that the chosen number of articles of 60 might be too scarce to represent the selected newspapers in a way to let us make very far-reaching and grounded conclusions regarding the nature of the reporting. Specifically, in respect of the comparison of discourse in Germany's conservative and liberal newspapers. Again, we view our study here rather as a basis for comparison as opposed to an introduction to a hypothesis.

8 Conclusion

In summation, the question of whether various discourses on AI exist across the board of differing political-economic systems is tackled in this paper. The largest national economic-political powerhouses and AI-contributors are chosen as representatives, namely the USA, China and Germany, and popular newspapers in these countries are selected to be analysed. The differences in political systems in relation to the differences in sentiment and SWOT analysis, while not sufficient to make grand extrapolations, do provide an insight and starting point to understand the link between the two. It becomes clear that, in line with our thesis, communist China's push for global technological dominance means highly positive journalism, free-press and -market USA is more neutral but still pro AI opportunities and strengths, while democratic, yet cautiously critical Germany is more aware of the threats and dangers AI poses. Opportunities for further research include, analysing if there is a correlation to how differing AI discourse affects a nation's ability to develop and innovate in the field. For example, what is the effect of China's extremely one-sided discourse on the future of AI in the country? And how will their decisions overflow into the world?

References

Analytics Insight. 2020. *Analytics insight predicts artificial intelligence market at US $53.2 billion in 2020, North America to lead*. https://www.analyticsinsight.net/analytics-insight-predicts-artificial-intelligence-market-us53-2-billion-north-america-lead/ (6 March, 2021).

Brennan, Allison. 2012. *The New York Times endorses Obama again*. https://politicalticker.blogs.cnn.com/2012/10/27/the-new-york-times-endorses-obama-again/ (6 March, 2021).

Bundesregierung. 2021. *Strategie Künstliche Intelligenz der Bundesregierung*. https://www.bmbf.de/files/Nationale_KI-Strategie.pdf (25 March, 2021).

Cave, Stephen, Kate Coughlan & Kanta Dihal. 2019. "Scary robots": Examining public responses to AI. In *Proceedings of the 2019 aaai/acm conference on ai, ethics, and society* (AIES '19), 331–337. Honolulu, HI, USA: Association for Computing Machinery. DOI: 10.1145/3306618.3314232.

Cheng, Evelyn. 2021. *China says it now has nearly 1 billion internet users*. https://www.cnbc.com/2021/02/04/china-says-it-now-has-nearly-1-billion-internet-users.html (10 March, 2021).

China Daily. 2021a. *About China Daily*. https://www.chinadaily.com.cn/english/static/aboutchinadaily.html (8 March, 2021).

China Daily. 2021b. *China Daily's print media*. https://www.chinadaily.com.cn/static_e/printmedia.html (8 March, 2021).

Ching-Hua Chuan, Su Yeon Cho, Wan-Hsiu Sunny Tsai. 2019. Framing artificial intelligence in American newspapers. *Spotlight 3: Empirical Perspectives*. 339–344. https://doi.org/10.1145/3306618.3314285.

David Broockman, Greg Ferenstein & Naresh Malhotra. 2017. Wealthy elites' policy preferences and economic inequality: The case of technology entrepreneurs. *Social Science Research Network*. DOI: 10.2139/SSRN.3028373.

Dominique Brossard, Dietram A. Scheufele. 2013. Science, new media, and the public. *Science* 339(6115). 40–41.

Fast, Ethan & Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence* 31(1). 963–969.

Fischer, Sarah, Carla Hustedt, Markus Overdiek & Ralph Müller-Eiselt. 2021. Wie Deutschland über Algorithmen schreibt. *Algorithmenethik*. https://algorithmenethik.de/2021/02/04/wie-deutschland-ueber-algorithmen-schreibt/.

Garvey, Colin & Chandler Maskal. 2020. Sentiment analysis of the news media on artificial intelligence does not support claims of negative bias against artificial intelligence. *Omics: A Journal of Integrative Biology* 24(5). 286–299.

International Monetary Fund. 2020. *United States at a glance*. https://www.imf.org/en/Countries/USA#countrydata (6 March, 2021).

Kelion, Leo. 2020. *EU reveals plan to regulate Big Tech*. https://www.bbc.com/news/technology-55318225 (6 March, 2021).

Lam, Taylor, Frank Li & Jeff Loucks. 2019. *China emerges as global tech, innovation leader*. https://deloitte.wsj.com/cio/2019/10/30/china-emerges-as-global-tech-innovation-leader/ (8 March, 2021).

Lee, David. 2019. *Artificial intelligence adoption gathering momentum in China*. https://www.chinadaily.com.cn/a/201911/20/WS5dd49966a310cf3e35578967.html (8 March, 2021).

Levitz, Eric. 2016. *A.G. Sulzberger vanquishes his cousins, becomes deputy publisher of the New York Times*. https://nymag.com/intelligencer/2016/10/a-g-sulzberger-becomes-deputy-publisher-of-new-york-times.html (6 March, 2021).

Lucey, Bill. 2010. *The New York Times: A chronology: 1851-2010*. https://web.archive.org/web/20170225020720/http://www.nysl.nysed.gov/nysnp/nytlucey.html (6 March, 2021).

Manjoo, Farhad. 2017. *Tech's frightful five: They've got us.* https://www.nytimes.com/2017/05/10/technology/techs-frightful-five-theyve-got-us.html (6 March, 2021).

Meyer, David. 2021. *Germany and France push forward with Big Tech crackdown, rather than waiting for EU-wide laws.* https://fortune.com/2021/01/19/germany-france-antitrust-content-moderation-eu-dsa-dma/ (6 March, 2021).

Ouchchy, Leila, Allen Coin & Veljko Dubljević. 2020. AI in the headlines: The portrayal of the ethical issues of artificial intelligence in the media. *AI & SOCIETY* 35(4). 927–936.

Pew Research Center. 2012. *In Changing News Landscape, Even Television is Vulnerable. Section 4: Demographics and Political Views of News Audiences.* https://www.pewresearch.org/politics/2012/09/27/section-4-demographics-and-political-views-of-news-audiences/ (6 March, 2021).

Roberts, Huw, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang & Luciano Floridi. 2020. The chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI&Society* 36(6115).

Schmidt, Karl Patterson. 2021. *United States.* https://www.britannica.com/place/United-States (6 March, 2021).

Song, Candice. 2021. *English levels in China: Quality of spoken english, signage, etc.* https://www.chinahighlights.com/travelguide/english-levels-in-china.htm (9 March, 2021).

The New York Times Company. 2021. *Our journalism.* https://www.nytco.com/our-journalism/ (6 March, 2021).

The White House. 2021. *Our government.* @online%7BRSF2020%7D.

United States Embassy. 2021. *The role of government in the economy.* https://usa.usembassy.de/etexts/oecon/chap6.html (6 March, 2021).

Westerheide, Fabian. 2020. *China - The first artificial intelligence superpower.* https://www.forbes.com/sites/cognitiveworld/2020/01/14/china-artificial-intelligence-superpower/?sh=798911322f05 (10 March, 2021).

Wikipedia. 2021a. *Frankfurter Allgemeine Zeitung.* https://de.wikipedia.org/wiki/Frankfurter_Allgemeine_Zeitung (6 March, 2021).

Wikipedia. 2021b. *Süddeutsche Zeitung.* https://de.wikipedia.org/wiki/S%C3%BCddeutsche_Zeitung (6 March, 2021).

Wikipedia. 2021c. *Zeit Online.* https://en.wikipedia.org/wiki/Die_Zeit#Zeit_Online (6 March, 2021).

World Press Freedom Index. 2020. *2020 World Press Freedom Index.* https://rsf.org/en/ranking (8 March, 2021).

Xu, Yixiang. 2019. *Meet China's AI competition: Germany's drive toward AI innovation needs sound policy and partnership with U.S.* https://www.aicgs.org/2019/03/meet-chinas-ai-competition-germanys-drive-toward-ai-innovation-needs-sound-policy-and-partnership-with-u-s/ (6 March, 2021).

Zeng, Jing, Chung-Hong Chan & Mike S Schäfer. 2020. Contested Chinese dreams of AI? Public discourse about artificial intelligence on wechat and people's daily Online. *Information, Communication & Society* 2020. 1–22. DOI: 10.1080/1369118X.2020.1776372.

Chapter 14

Ethical concerns and AI: Analysing British news articles about Alexa

Kim Targan, Clara Matheis & Kristof Engelhardt

The recent successes in the field of Artificial Intelligence (AI) have increased the awareness of everyday AI applications. This has led to discussions about so-called smart devices, commonly labeled as AI. While ethics of AI does receive some attention, ethical discussions in relation to commonly used AI-based services are still limited. In this paper we review how ethical concerns regarding the voice assistant service ”Alexa” are portrayed on UK news websites. The review of articles enables us to draw conclusions about the circumscribed public awareness about ethics in relation to everyday “intelligent devices”. The analysis reveals that the main focus is placed on concerns about the role of human beings in data review (data security) and development of services (bias). The articles express recommendations which indicate that voice assistant services in general are seen as beneficial, while there are still concerns about the current implementation and handling. Finally, the criticism suggests possibilities of future developments and advises consumers to adapt their usage habits.

Keywords: Artificial intelligence | News articles | Ethical questions | Voice assistant | Amazon

1 Introduction

The recent successes in the fields of voice recognition and Question-Answering have enabled a steep incline in the commercial sale of virtual assistants. With around three billion used voice assistants, they already play a central part in the everyday life of many people (Gayle 2020). In the last years a lot of privacy concerns in relation to so-called smart speakers were expressed, nonetheless the number of users is increasing continuously: every 9th household in Germany uses a smart speaker, whereas in the UK it is already every 5th (Kinsella 2019). This discrepancy raises the question of how ethical concerns with voice assistants are really discussed in the public discourse and what consequences for the consumers are addressed. This paper aims to approach these questions through analysing news reports. In the following we will analyse the portrayal of ethical concerns in UK news articles about AI based voice assistants by taking the example of Alexa. We will start by providing a short introduction to "Ethics of AI" and addressing the process of data selection and analytical methods employed. The results of our primarily qualitative analysis will be discussed in Section 4 and can be divided into four main fields of interest. First of all, we are going to analyse which ethical concerns are expressed to delineate the major concerns about Alexa that are present in media discourse. The second field of interest can be summarized as "suggestions" and is concerned with action recommendations. Thirdly, we are going to take a look at the tone of these articles - specifically, the way that Alexa is portrayed and the use of emotive language. In this section we will discuss details about the relation between the tone of the article and ethical concerns. Lastly, the analysis will tie back to the overarching topic of AI, focussing on the classification of Alexa as AI as well as the general portrayal of AI.

To motivate the choice of our example, we will start by providing a short introduction to Alexa. Alexa is the name of Amazon's "cloud-based voice service" (Kim 2018) and it is generally used to denote the virtual assistant service offered by the company. It can operate on various devices, however, it is most often used as a service embedded in Amazon Echo smart speakers. We decided to focus on Alexa, as Echo speakers are used by customers deliberately for voice assistance. Thus, in contrast to Apple's "Siri" (used on iPhones) or "Google Assistant", Alexa is not just used as a feature of a pre-existing device (Kinsella 2019). Users explicitly buy the device for using the voice assistant. Thus, the choice of Alexa allows for a more specific analysis of ethical concerns around the use of voice assistant services. News articles frequently classify Alexa as an AI or AI-based speaker (for references see 4.3.1), thus the discussion can be located in the public discourse about Artificial Intelligence. Implementation and architectural details will be further discussed in Section 4.3, nevertheless a short introduction can be offered: the speech recognition and natural language generation at the core of Alexa is based on deep neural networks which are trained in both supervised and unsupervised manner. Supervised workflows include human reviews and labeling of the underlying data to provide a 'ground truth' against which to measure the models performance, while unsupervised learning is performed on unlabeled data. The main Artificial Intelligence component of Alexa is online unsupervised learning - the way “corrections [are made] based on context clues automatically” during user interaction (Horowitz 2020). Finally, by using Alexa, an increasing number of people are interacting with a device partly based on AI. Thus, the discussion about ethics of AI and ethical concerns in relation to Alexa is becoming increasingly relevant.

2 Ethical concerns and AI

In this section, we will shortly motivate our focus on ethics and give an overview on the field of “Ethics of AI”. With the current pace in which new inventions in the field of AI are popping up and delivering solutions for long explored problems, the appeal grows to just try them out and see what happens. In a lot of cases this actually works quite well: while the learning process and features analysed may not be interpretable by humans, the results are often impressive. However, it can happen that an AI performs well in test phases, but only on what is being tested. One does not know exactly which biases are learned during training based on the data. In addition, if there is limited concrete legislative regulation, it is not essential for companies to include ethical considerations into their project as there is little financial profit linked to it. Thus, ethical questions can fade into the background. We think this makes it even more necessary to include the central topic of ethics into this fast growing field of research and questions about the underlying algorithms, the biases replicated and finally, the possibility of ethical use and creation of AI based technologies.

The field ”Ethics of AI” belongs to the domain of Applied Ethics and is comparatively young. It addresses questions concerning how AI systems should be used, what they are supposed to do and what risks are involved in their use. Nonetheless, current discussions often focus mainly on several specified fields of ethical concerns (Müller 2020: 1). Media coverage of Ethics of AI demonstrates this focus and can thus be categorized along the dimension of concern type. For the purpose of our research, we have focussed upon two main groups of ethical issues that are associated with the use and design of AI systems: ”privacy & surveillance” (Müller 2020: 1) as well as ”bias in decision systems” (Müller 2020: 1). On this basis we will summarize the ethical questions that may arise depending on the type of use and the design choices of AI.

3 Data and Methods

3.1 Data

We have chosen the medium news articles as they represent current discussions and the concerns of the public quite well. In contrast to more thematically specific journals, the most common online newspapers have a large readership. As we want to reflect the public discourse, this diverse readership gets us as close to the general non-expert public as possible. In this paper we are going to focus on articles from the UK news-websites that received the highest ranking in a popularity score of the web-based content reader Feedspot (Agarwal 2021). We chose this ranking because it does not only rank newspapers by their circulation like rankings we found on Wikipedia, but also includes their online presence through evaluating social media followers and engagements. In our data selection we used this website rating as well as further search engine rankings as a proxy for relevance of the articles to public discourse. However, while we did try to cross-reference search query rankings this will likely be dependent on several factors we were not able to control for. A list of the resulting article selection is included in the appendix.

As mentioned above, we chose to focus our analysis on two fields of ethical concerns as queries focussed solely on 'Alexa' and 'ethics' only provided limited results. The article corpus was established through two main query processes. First, we used a Google query to determine an appropriate query that resulted in suggestions for all nine news websites with the highest ranking. This query finally consisted of "*website* Alexa [privacy,bias]". Secondly, the nine news websites were queried for "Alexa [privacy,bias]" directly. To limit the scope of our research and focus on the discussions that are closest and most relevant to the current version of Alexa, we limited the time frame to be analysed and took only articles issued after 1 January 2019 into account. Besides, we could only include articles published before the start of our data selection in December 2020. From the resulting article database we excluded all articles that did not focus on Alexa or ethics related topics in the article itself. We then picked the three highest ranked articles of each news website. The most important factor for finalising our article selection was the result of the website query. Due to our selection criteria, there were several news websites that listed only three or less articles. Thus, we did include fewer articles for some news websites due to limited resources ("The Sun", "Daily Express", "Metro"). Our resulting corpus is limited to UK news website articles from January 2019 to December 2020 and consists of twenty-two articles.

3.2 Methods

For our analysis we used both quantitative and qualitative methods. Specifically, the paper comprises an evaluation of coding of the newspaper articles with different schemes for language and tone, ethical concerns, suggestions and relation to AI. As coding is often referred to as tagging, in the following we will be using code and tag interchangeably. The initial starting point for our investigations was characterized by quantitative analyses of the coded data to identify phenomena and get insights into correlations between tags. As a second step, we verified these observations of patterns of occurrence by looking at the codes directly and used the insights to guide our attention towards specific codes to draw conclusions about the nature of the discourse on ethics. Our tagset was defined to reveal not only content information about the articles, but also the language use. The content tagset consisted of three codes with subcodes: ethical concerns, suggestions and AI related tags. For analysing the tone of the article we included a tagset with different kinds of portrayals of Alexa and one consisting of types of language use, with subcodes including emotive language and expert justifications. A more detailed overview of our codeset can be found in the appendix. The coding process itself was divided into several stages. A first scanning and open coding of the articles revealed the concerns and suggestions repeatedly expressed in the articles and allowed for establishing a comprehensive tagset (use of open coding as referenced in Flick 2013: 270). Secondly, all data from the first step was deleted and all articles were tagged once. Lastly, all articles were cross-tagged by at least one other person to ensure that codes were applied consistently. Code meaning and appropriate use are discussed repeatedly during the tagging process and special care was taken to ensure assigning all relevant codes to a passage, thus many passages are coded to three or more codes.

3.3 Limitations of the approach

The data that lies at the heart of this evaluation is limited to the discourse on UK news websites. Thus, it cannot represent the discourse in other countries. Further comparison of the discourse between countries, including the USA where there is currently the largest recess of Echo devices would definitely be worthwhile and might reveal differences in the handling of these ethical concerns. Another interesting topic for further analysis which goes beyond the scope of this discussion would be an evaluation of differences between the news websites. The style of reporting and ground covered varies a lot by virtue of the type of news website it is published on. In our analysis, we included a sample of different types of websites based on the above mentioned popularity ranking to counteract biases introduced through the focus on a specific style of reporting. Thus, this analysis aims to cover a larger sample of styles and does not focus on the differences between sub-spaces of news discourse. Besides, the article selection is certainly not exhaustive even though there are surprisingly few articles on ethical concerns regarding Alexa. On the basis of this restricting factor, quantitative findings and analyses cannot approximate the complete discourse in UK news article. This is additionally heightened by the fact that especially the coding scheme for language use included various more subjective codes such as the code "Emotions". To eliminate "intracoder" biases we did crosscode all articles, but the evaluations on these tags are nonetheless inherently subjective. Finally, a lot of the chosen articles actually respond to revelations about data security. In response to the resulting public outcry, Amazon has adopted new privacy control features and released new products. Thus, both Amazon's privacy policy and products have changed within the considered time period.

4 Results

4.1 Ethics and AI

The first step of analyzing the discussion about ethical concerns is to investigate how and if ethics is mentioned in the articles. We searched the whole text corpus for instances and deflected forms of the word "ethics" in order to see if the topic is explicitly mentioned. The query yielded no results, therefore it can be concluded that ethical questions are not explicitly marked as such in our article selection. Therefore, the discussion of ethical concerns is for the most part implicit. The articles do not focus on why issues are concerning on a philosophical level, instead many of the news articles react to revelations of Amazon's data handling or the release of an UN report (West et al. 2019), without addressing the topic of data security or bias as a whole. The societal implications of problematic practices are rarely addressed. Although we encountered no explicit discussion of "ethics", a variety of ethical questions and concerns are discussed implicitly.

4.1.1 Ethical concerns

In the following, we will delineate which ethical concerns are mentioned in the articles. Furthermore, we will analyse in how many articles each concern is mentioned to determine the importance ascribed to the concerns by UK news websites. On the basis of our coded corpus we extracted quantitative measurements of the occurrence of the different codes. For the analysis of the coding scheme "Ethical concerns" we split the data into articles about privacy, 14 in total, and articles about bias, 7 in total. Additionally, a Telegraph article (Johnston 2020) tackled both topics, therefore we included it in the analysis of both topics in the following. An overview over the numerical distribution of codes across articles can be seen in Table 1.

Table 1: Appearance of codes for ethical concerns in articles

Code	Total*	Data security articles	Bias articles
Bias	2	0	1
Bias: Race	2	0	1
Bias: Gender	7	1	6
Data security	13	12	0
Data security: Privacy	13	12	0
Data security: Surveillance	7	7	0
Transparency	12	10	1
Transparency: Privacy policy	10	10	0
Juridical guidelines	8	6	1
Technical mistakes	9	6	2
Human reviewers	10	10	0

*Includes the counts for one article that tackles both topics, thus the total can exceed the sum of the ”Data security” and ”Bias” column

In the privacy corpus 13 out of 15 articles explicitly mention data security or privacy as ethical concerns. A special focus seems to be on the issue of human reviewers, who transcribe Alexa recordings in order to improve the voice recognition and other systems involved to allow for supervised learning. It is seen as inherently problematic that other humans listen to private recordings, while in comparison, automated interpretation by algorithms seems to not be a central issue. An article by the news website of the Independent even wrote: "Amazon previously denied that its Echo devices were used to spy on people but [...] admitted that employees listen to customer voice recordings" (Cuthbertson 2019a), implying that it can mainly be called spying if humans directly listen to the Alexa recordings. Only five of the privacy articles elaborate on privacy concerns related to the general use of Alexa data to commercial or administrative ends (E.g. Johnston 2020) or the potential threat of hackers accessing very personal information (E.g. BBC News 2020). Besides not wanting people to hear private conversations in general, some articles that deal with human reviewers portray the fear of being identifiable or of involuntary sharing sensitive information like bank accounts login data and private conversations. A Mirror article writes "According to a report in The Sun, Amazon Staff overheard private moments including family rows, money and health discussions - and even couples having sex. [...] A former employee told the newspaper that staff were told to focus on Alexa commands, but that it was 'impossible not to hear other things going on'" (Curtis 2019b). This is portrayed as especially concerning in connection to another ethical concern mentioned in nearly half of the articles, namely technical mistakes. Alexa is heavily criticised for recording even in cases where the wake word "Alexa" is not even mentioned (E.g. Lynskey 2019). This led to the recording of very private conversations of customers, which subsequently could be reviewed by humans.

Another big concern is transparency, for which the focus is again placed on the issue of human reviewers. Nearly all of the eleven articles mentioning transparency issues in the context of privacy are concerned about Amazon's privacy policy. The articles criticize that the policy does not clearly expose that the recordings of Alexa could be subject to human review. In contrast, only two articles are concerned about the non-transparency regarding the use of data recordings in general.

For the subset of bias articles, all but one article heavily focus on gender bias, while race and general biases are rarely mentioned with just one or two occurrences in articles per tag. The reason for the greater attention to gender bias over other biases might be due to Alexa's voice being perceived as female. In addition, concerns about the use of a female voice received strong attention in reaction to an UN report (West et al. 2019), which reported ethical concerns with regard to gender biases in virtual assistants (more on the report in 4.2.1).

The main ethical concern regarding gender bias is the perpetuation and reinforcement of the stereotype that women are subservient. Most of the articles see the problem in the types of responses of Alexa. For example a Guardian article writes "the often submissive and flirty responses offered by the systems to many queries - including outright abusive ones - reinforce ideas of women as subservient" (As cited in Rawlinson 2019 in reference to the UN report West et al. 2019).

Interestingly, only two articles, one from the Daily Express and one from the Telegraph, mention racial bias, each in one sentence. The ethical concerns however are not addressed specifically in regards to Alexa, rather in relation to algorithms that use AI in general. They give the example of Black people being discriminated against by algorithms in healthcare decisions (Fisch 2020) or in algorithms measuring recidivism used for bail decisions (Johnston 2020). The reason for the apparent lack of discussion of racial bias might be, that racial discrimination is more subtle in data consisting only of audio recordings and probably mainly that racism is still often swept under the table.

These two articles are also the only ones that address bias in general. They are concerned with data being inherently biased which can lead to discrimination and perpetuation of biases that already exist. The Express cites the director of decisioning at Pegasystems who said: "AI is as biased as the data used to create it. Even if designers have the best intentions, errors may creep in through the selection of biased data for machine learning models, as well as prejudice and assumptions in build-in logic" (Fisch 2020). To us it is interesting that this topic did receive comparable little attention, since we think it is a central underlying ethical problem. If this problem is not addressed, this will lead to discriminatory biases in the first place.

Juridical guidelines, transparency, data security, privacy and human reviewers did not receive much attention in the articles in the bias corpus.

All concerns which occurred within the articles, were responses to either big ethics investigations by independent organisations or to dramatic reveals of bad practises. All the articles can be traced back to a few of those events, triggering the release of various articles concerning the topic in a short period of time. We did not find articles that describe a general problem with ethics concerning Alexa which is not connected to such an event in response to our search query. In general, the addressed issues are mostly limited to concerns revealed by such big events and only few articles write about Alexa in a more extensive manner (Cf. Johnston 2020). One major event was a statement by Amazon in April 2019, admitting that some voice recordings were being manually reviewed to improve voice recognition (Cuthbertson 2019b). We did not include any articles before this point, however it started a series of critical reports at once (E.g. BBC News 2019, Daily Mail 2019b, Cuthbertson 2019b). This outcry was followed by a change of privacy settings in August, which allowed costumers to stop the usage of their data for the improvement of Amazon Services (Curtis 2019b). The option of preventing your recordings from being manually reviewed, was afterwards often referred to as an ”opt-out" option, however it did not change the overall critical perception.

4.1.2 Suggestions

In the face of ethical concerns, most of the articles have suggestions about what needs to happen moving forward. These suggestions can be divided into subsets based on the type of agents they are directed at: the Alexa users, the company Amazon and thirdly governmental officials. The different forms of advice directed at the users range from filing juridical complaints against the company Amazon, to regulatory user control or no required action at all. Demands of action from Amazon place the handling of the ethical concerns in the hands of the company and call for specific action plans. Lastly, some articles include demands of action from governmental agents, calling for legal (AI) regulation.

An analysis of the numerical distribution of suggestions across articles revealed that most articles demanded action of Amazon, specifically 17 of 22 articles. We captured these demands with the code "Responsibility of Amazon" and related subcodes - thus, these subcodes provide an insight into the nature of the proposals. The demand expressed most often towards Amazon is a call for more diversity in its staff, followed by urges to invest in data protection and calls to delete recordings. As analyzed below, the distribution of subcodes is dependent on the ethical concerns expressed in the article. Suggestions directed at the user, captured with the code "user control", were expressed in half of the articles. Several of them proposed to use the "opt-out" options provided by Amazon to control handling of their data. Finally, six articles proposed that it would be desirable to develop a more ethical system instead. Demands of action from governmental officials are least common, only two articles criticized the absence of legal regulations.

As a part of our qualitative analysis, we did analyze the suggestions as a proxy to assess the seriousness that is attributed to ethical concerns of Alexa. Explicit calls for legal prohibition and refraining from using Alexa convey the message that ethical concerns are to be taken very seriously. Therefore, these are located on the more serious side of the spectrum of concerns, while simply using Amazon's "opt-out" option is located towards the opposite side. Extreme suggestions on both sides of the spectrum are rare. On the one hand, no articles suggested using Alexa products as assistive devices without proposing further regulatory user control or demanding action of Amazon. However, the proposal made to users most frequently goes in a similar direction. It is suggested to use Amazon's "opt-out" option to disallow Amazon to use any recorded data for improving their services. The frequent occurrence of this suggestion implies trust in the option or at least satisfaction with the setting. This indicates that the concerns with Alexa and the company Amazon are not perceived as majorly threatening.

On the other hand, there are no demands for legal prohibition of Alexa and only two suggestions to use governmental regulations to mediate ethical issues (Lynskey 2019, Johnston 2020). However, one Guardian article (Lynskey 2019) explicitly mentions that legal regulation is the most important measure to be taken. The Gizmodo editor Adam Clark Estes is quoted to say: “I hate to be dramatic, but I don’t think we’re ever going to feel safe from their data-collection practices. Government regulation is the only thing that is going to halt more damage” (as cited in (Lynskey 2019)). It is definitely important to note that only one article articulates this message, while all others do not address the possibility of governmental action. Besides, there is only one other article that uses the word “government”. Thus, overall, actions are discussed as a main possibility for customers and company, while legislative restrictions are not considered. Another suggestion that can be interpreted as taking concerns very seriously, is to stop using Alexa based devices altogether. This is expressed by the Guardian article mentioned above (Cf. Lynskey 2019) and one other article (Winder 2020). These suggestions are formulated insistently, sometimes even implicitly criticising Alexa users, as can be seen in the Guardian article (Lynskey 2019): “If you still have an Alexa or any other voice assistant in your home, you were warned”. This passage and the one quoted above, illustrate the important insight that many of the more extreme suggestions are expressed through quotations, however this will be discussed in more depth in 4.2.1. Finally, most suggestions comprised less extreme measures.

4.1.3 Correlation of Suggestions with Ethical Concerns

Further quantitative as well as qualitative analysis reveals that the suggestions proposed in an article correlate with the expression of specific ethical concerns. For many of the suggestion and concern pairs this relationship is meaningful: suggestions occur in the context of specific concerns as the proposals are made in response to the ethical concern. As outlined above, the article corpus can be divided into two subcorpora: one with a main focus on bias and one focussed on data security. Figure 2 shows the numerical distribution of suggestion codes across articles for the different subcorpora. For this analysis the absolute value of occurrence within an article is not investigated. The distribution of codes indicates that articles concerned about the issue of bias, propose four main remedies. Most importantly there is a call for Amazon to diversify its staff. Secondly, there seems to be faith in the possibility of creating a more ethical voice assistant, specifically, a voice assistant that does not have a default female voice. Besides, there are some demands to increase transparency directed towards Amazon. Overall, the responsibility to address biases inherent in the system is delegated to Amazon. Only one article suggests implementing legal regulation. Summing up, it can be said that the ethical concern bias is believed to be best addressed by either more diversity in Amazon’s developers or the creation of a voice assistant with a gender neutral voice. There is little discussion about the inherent bias in training data and thus the content of answers (of even a gender neutral voice) which are not simply mitigated via more staff diversity. The recommendations of articles concerned with data security are more varied. In the context of privacy concerns, suggestions are firstly directed at the user. Users are informed about Amazon's privacy settings, most importantly the option to "opt-out" of the use of the recordings for improving Amazon's service. Other proposals include the advice to switch the Echo device (on which Alexa operates) off whenever one is concerned about privacy. Finally, two articles suggest not to use Alexa. Additionally, the articles demand action from Amazon to protect the data collected or delete the recordings. Finally, there are some articles that see promising possibilities for legal regulations as well as conjoint juridical complaints of the users. Overall, articles on data security seem to follow the underlying sentiment that regulatory usage strategies on the side of the customer are an important if not the main remedy to mediate privacy concerns.

Table 2: Appearence of codes for suggestions in articles

Tag	Total*	Data security articles	Bias articles
Assistive device	1	1	0
Develop ethical tool instead	6	1	5
Judicial complaints	3	3	0
Legal regulation	2	1	0
Responsibility of amazon (RA)	2	0	1
- RA: data protection	3	3	0
- RA: deletion of recordings	4	4	0
- RA: diversity	7	0	7
- RA: transparency	5	3	2
User control	5	5	0

*Includes the counts for one article that tackles both topics, thus the total can exceed the sum of the "Data security" and "Bias" column

4.1.4 Futuristic outlook

For our qualitative analysis we did not only examine current ethical issues which are mentioned in the articles, but also considered how the future of Alexa and Artificial Intelligence are portrayed by using the code "futuristic outlooks". There are two major types of outlooks. Firstly, there are negative outlooks which foresee a worsening of current privacy issues. The second group paints a future with possible solutions for the current ethical questions. There are articles that include both categories (E.g. Lynskey 2019). The Guardian article cites different sources and finally ends on a rather positive note expressing hope to find solutions without explicitly giving examples. Other articles are very specific in their suggestions on how to build a more positive future, they demand new privacy features and fully user-controlled voice assistants for a safer handling of data or gender-neutral voice assistants to tackle problems concerning gender bias. Three articles (Cf. Dawson 2019, Daily Mail 2019a, Rawlinson 2019) cite the UN-Report (West et al. 2019), which suggests to "programme digital assistants to discourage gender-based insults". However, this is an exception and only shortly mentioned. Most articles refer to new inventions improving safety, rather than restricting any existing features or the assistants themselves. In the negative category, one article refers to implanted microchips, while others pointed out the issue of using Alexa data for criminal investigations, to just name a few. However, there are also some positive opportunities pointed out. One Guardian article (Rawlinson 2019) refers to the UN-report (West et al. 2019), which shows that the "recent introduction of such technology provided the opportunity to develop less damaging norms in its application". In general, very dystopian futuristic outlooks are seldom and mostly expressed in quotes. These extremely negative visions are sometimes portrayed as inevitable by the interviewee, nonetheless these quotes do not represent the overall sentiment of the articles, as further discussed in Section 4.2.5. The positive futuristic outlooks seem to propose future technical inventions, rather than a regulatory restrictive handling of the current situation.

4.2 Tone and language

A second field of interest when analysing journalistic articles is the use of language and tone of the article. In the following, we will address several questions we have investigated on the basis of the numerical distribution of our codes as well as a qualitative analysis of the data. We will start by analyzing the language use by examining the use of quotations and the expression of emotions. On this basis, we will evaluate the overall attitude towards Alexa.

4.2.1 Presentation of opinions

Quotations are one topic of concern in the domain of language use. News articles are generally expected to objectively delineate different stances and positions towards the topic at hand. This might be achieved by presenting both critical as well as positive stances in the form of quotations. Thus, an important question is by whom the concerns and suggestions are expressed. As mentioned above, it must be noted that radical suggestions and concerns are mostly expressed through quoting people depicted as experts. Commonly these experts are either affiliated with technology media, so-called cyber security institutions or former Amazon employees. However, a thorough examination of the agents who are consulted or quoted reveals different backgrounds of experience. On the one hand, articles are referring to professionals in the specific field, this can be observed in the case of juridical suggestions which are outlined through quoting lawyers (Murphy 2019). Further examples include the field of bias, where both concerns and suggestions are most often expressed through referring to an UN report (Cf. West et al. 2019) on gender biases perpetuated by AI voice assistants (Chowdhury 2019, Harris & Best 2019, Daily Mail 2019a, Dawson 2019). The report aims to highlight gender biases inherent in products in the technology sector, while also focussing on gender gaps in digital skills education. In addition to a general critique, the report suggests solutions for closing this digital skills gap (West et al. 2019: 37). Further quotations of agents with professional knowledge on the subject matter include security researchers (Cf. Daily Mail 2019b) and references to the so-called ”privacy watchdog”, the “supervisory authority for the company [Amazon]” (Kelion 2019). Finally, a minor subfield of suggestions discussed in the context of Alexa, is concerned with weighing safety benefits of AI systems against privacy concerns. In this context the article (Johnston 2020) referred to the statement of a “Met commissioner” who believes in the necessity of AI systems for safety, this is however not an expression of expertise but rather an opinion. The subgroup of agents, who are often interviewed are people with a large interest in the field. Nonetheless, they are neither clearly depicted as experts, nor necessarily Alexa users. This includes authors of books on (dystopian) futuristic visions of data security and privacy advocacy groups (Curtis 2019a). Statements made by these agents with a strong interest in the field often express very explicit suggestions and concerns. For example a Guardian article cites the author of the novel ”Zed”, Joanna Kavenna: ”[I]n Zed, the tech monopoly, Beetle, is omnipresent and unaccountable. ’The democratic idea is that we’re meant to have transparent corporations and governments, while people have privacy,’ she says” (Lynskey 2019).

On the other hand, articles often focus on agents with user experience. Especially in the realm of user-control many articles simply present anecdotal cases of prominent users that propose specific regulations on the use of Alexa. Repeatedly articles choose agents with former or current Amazon affiliation which might suggest their expert status. Nonetheless, in several cases the employees experience is just usage based with no special expertise on Alexa. Suggestions from former Amazon employees are reported in different articles (Cf. Keach 2020, Parsons 2020, Lynskey 2019) and often include very specific proposals, such as referring to "Robert Frederick, a former Amazon executive, [who] said he turns his Alexa off for privacy" during private conversations (E.g. Parsons 2020). Additionally, there are several examples of articles simply referring to a (cyber-security) expert, sometimes without providing further information (Fisch 2020). As an example, one person introduced as a privacy expert is reported to suggest that she does not use Alexa in the bedroom (Cuthbertson 2019a).

While the spectrum of agents that are referenced above mostly take skeptical stances towards Alexa, more than half of the articles also express a contrary opinion through quoting Amazon's official spokesperson. These statements often highlight that Amazon is dedicated towards data security and outline the privacy setting options customers have. Many articles even end with uncommented quotes of Amazon's spokesperson. Thus, the final impression of these articles is shaped by Amazon's self-portrayal: "[w]e take the security and privacy of our customers' personal information seriously" (as cited in (Cuthbertson 2019b)). Depending on the general content of the articles, this either serves as a counterpoint to the expression of concerns included earlier on (E.g. Kelion 2019, Curtis 2019a) or suggests a positive stance of the article towards Amazon's handling of ethical issues (Cf. Parsons 2019). So all in all, quotations of Amazon diversify the range of attitudes expressed towards Alexa and often introduce the possibility of "User-control" in the article (E.g. Parsons 2019).

4.2.2 Anthropomorphisation of Alexa

One striking finding when analyzing the portrayal of Alexa, is the coocurrence between the use of humanizing language and the expression of emotions. In articles where Alexa is referred to as human-like, for example by using "she/her" pronouns, the code "Fear" occurs quite often. The emotive language indicates that there is an increased fear of humanised technical devices. Human-like presentation might increase the perception that Alexa is an autonomous agent and thus lead to feelings of threat. Portraying AI as human, using "she/her" pronouns or in general assigning a gender (E.g. Harris & Best 2019) is a common practice. In addition, subservient and servile terms are frequently used in relation to Alexa. One article describes Alexa, by stating "I have a new servant who wakes me up in the morning" (Johnston 2020). Combined with humanizing vocabulary this sentence conveys the image of a submissive female servant. One reason for the use of "she/her" pronouns might be that Alexa is a name that is perceived as female. Plus, the voice that is used is high-pitched and generally perceived as female. Thus, the gender perception is probably implicitly swayed by Amazon's name and also voice choice. In addition to the attribution of human roles, the articles describe "Alexa's actions" in terms of human behaviour. This intensifies the feeling of technical surveillance, by depicting the threat of a human-like observer with expressions like "eavesdropping" or stating that Alexa "know[s]" (Cuthbertson 2019a). However, some articles also criticise the constant anthropomorphisation of Alexa (E.g. Petter 2019), for changing the users interaction with Alexa. Nevertheless, in general the news articles frequently use humanising vocabulary in relation to Alexa.

4.2.3 Emotive language use

The news articles use emotive language to report about Alexa. There are a few exceptions, however more than half of the articles express fear in relation to Alexa. As an example for fear expressed regarding privacy concerns a BBC article writes: "A common fear is that smart speakers are secretly recording everything that is said in the home" (BBC News 2019). Also fear is often expressed in relation to possibilities of (mis)use of the recorded data (E.g. BBC News 2019, Lynskey 2019, Daily Mail 2019b). Nevertheless, the expression of fear is often followed by suggestions which might weaken the perception of threat. In several articles, fearful passages are surrounded by Amazon privacy policy statements (BBC News 2020, Cuthbertson 2019b, Curtis 2019a), suggesting that this fear is unjustified (as discussed in 4.2.1). One specifically noticeable example of this combination with fear can be found in the article (Curtis 2019a). After in-depth elaboration on problems of voice recordings of children, which children's rights organisations harshly criticised (Curtis 2019a), a Mirror article ended with an opposing Amazon statement claiming to conform to child protection acts. Another emotion quite frequently used is "doubt". It is used in most contexts to question a relatively positive attitude towards Alexa. Doubt is mainly expressed by people who are interviewed rather than the authors themselves. The expression of "fear" and "doubt" is often correlating, as they can play into each other. However, the third most frequently coded emotion "euphoria" also co-occurs with "fear" in some articles. Often euphoric language is followed by a critique of this positive stance. To provide an example: "Although the baseline benefits of AI are indisputable, the cutting edge technology contains a dangerous caveat - inherent bias" (Fisch 2020) (Cf. Johnston 2020, Dawson 2019). Thus, articles do express euphoria or at least a positive stance towards the general idea of voice assistants but nonetheless do not endorse the specific example of "Alexa".

4.2.4 Emotive expression of ethical concerns

Connecting those emotions with ethical concerns provides a clearer picture of why emotive language is used and what it is supposed to convey.

What catches one's eye at first, is the correlation of the emotions of "Fear" and "Doubt" with ethical concerns of data security, privacy and surveillance. When data concerns are addressed, emotions of doubt towards companies or the fear of complete surveillance are often at play and help to intensify concerns. Both of these emotions also occur in the context of transparency concerns in general and in concerns relating to Amazon's privacy policy. Besides, the emotions "euphoria" and "content" do occur in articles about data security as well. However, as mentioned above, positive opinions and emotions are for the most part directly followed by expressions of concern about the different ethical issues, and thus do not outweigh the concerns. Finally, the articles present different reactions to questions about data security and bias which is reflected in the high cooccurrences with positive as well as negative emotions. One last very interesting finding is that there is no expression of content and euphoria in articles where juridical complaints are mentioned. In comparison to other suggestions, these complaints are most explicitly stated as demands. These articles portray that topic in a critical manner without relying on emotive language, especially not positive emotions.

Finally, most articles use a lot of emotions to point out and intensify the concern about ethical issues, however they do not represent very strong unilateral opinions (for exceptions see 4.2.5) and are depicting, as well as using, emotive language for both sides of argumentation.

4.2.5 Attitude towards Alexa

On the basis of the discussion of tone and language use provided above we can now move on to the general portrayal of Alexa. An analysis of the coded data revealed that the articles mostly express negative attitudes towards Alexa. We used the portrayal of Alexa as well as the expressed emotions to determine the overall tone of the article. Additionally, the code "Criticism of Alexa" sheds light onto the overall attitude towards Alexa. Several articles criticise Amazon for allowing gender biases in their products (E.g. Rawlinson 2019), not having diverse development teams and using a default female voice for Alexa (Dawson 2019). Sometimes the focus of criticism is on specific design features in Alexa's composition and development, rather than the device itself. Subcodes of "Portrayal of Alexa" range from positive to negative, but also include a code for explicit connections to danger and threat ("Dangerous"). Both "Negative" and "Dangerous" portrayals can be observed equally often and mainly both occur in the same articles. Analysing the cooccurrence of the attitudes expressed, one can see that the code "Positive" often appeared in the same articles as "Negative" and even together with "Dangerous". Therefore, for many articles, the language and the tone are quite diverse, even throughout the article, conveying positive as well as negative attitudes. The expression of mixed attitudes can be ascribed to the frequent use of quotations or references to "expert" opinions. News articles often aim to seem objective by expressing a diverse range of opinions. This might be a reason why the articles present several opinions towards Alexa. A qualitative analysis of the expression of "positive attitudes" shows that these mainly comment on the helpful features of Alexa such as: "If you ask Alexa why your chest hurts, it will [...] give you an answer, most probably taken from the NHS website. In principle, this is great." (Johnston 2020). However, the expressed benefits and the thereby positive portrayal of Alexa is generally followed by an elaboration on ethical concerns and thus criticism of Alexa (E.g. Johnston 2020, Parsons 2020, Dawson 2019). There is one exception, namely a Metro article which advertises for an Amazon Echo product with a "new focus on privacy" (Parsons 2019). The article expresses content about the measures Amazon has taken to ensure privacy. On the contrary, two articles share a remarkably negative perception by describing Alexa as "colonising the user's home" (Lynskey 2019) and analysing several different flaws (Johnston 2020).

Summing up, the articles express a range of stances, nevertheless the general attitude towards Alexa is rather negative. However, it is common for articles not to end on a bad note. This is achieved by presenting possible solutions, such as proposing gender-neutral voices against gender-biases (Rawlinson 2019) or stating that it is possible to create a voice assistant which will not interfere with privacy (Lynskey 2019). This presents problems as solvable and advertises for further inventions rather than restrictions.

4.3 Alexa and AI

Tying back to the overarching topic of Artificial Intelligence, this section will analyse the portrayal of AI in the news articles. In order to remain within the scope of this chapter, we are going to focus on the classification of Alexa as AI of itself and in relation to different thematic codes. This can be achieved through a quantitative analysis, correlating the classification of Alexa with other code sets. Nevertheless, the analysed corpus only consists of twenty-two articles, thus there are limitations to the interpretability of these analyses.

4.3.1 Classification of Alexa

The first and most important question is if Alexa is labelled as AI. The article corpora can be divided into two subsets based on the way Alexa is described. Out of the twenty-two articles only ten explicitly labeled Alexa as AI (AI-corpora) and twelve do not mention any link to AI. Articles that do not label Alexa as AI often use the term "smart speaker" to refer to Alexa. Interestingly, more articles that focus on the concern of bias explicitly label Alexa as an "AI-bot" or "AI voice assistant", than articles focussing on data security. This is due to the fact that most of the articles concerned with bias are written in response to the UN-report West et al. 2019 which explicitly talks about "AI voice assistants" (West et al. 2019: 5). A quote from this report that appears frequently conveys that the digital assistant is based on "AI systems" (As cited in Chowdhury 2019, Harris & Best 2019, Dawson 2019). Response articles also often adopt expressions such as "AI-powered virtual assistant" (Dawson 2019). Aside from references to the UN report, Alexa is called an "AI speaker" (Harris & Best 2019), "intelligent digital assistant" (Petter 2019) or even "Artificial intelligence-based smart speaker system" (Chowdhury 2019). Finally, an article of the Independent (Cuthbertson 2019b) also refers to the "Alexa AI" and the Express article (Fisch 2020) mentions Alexa as an an example of "everyday AI", while one Telegraph article (Johnston 2020) refers to "Artificial Intelligence in our homes".

To discuss the accuracy of these labels, one has to discuss the actual technical details of Alexa. Alexa is a software agent that is embedded into devices such as Amazon's Echo products. The Echo device itself records speech which is interpreted by the cloud-based Alexa Voice Services. Thus, Alexa is based on automatic speech recognition, wake-word detection and an audio-to-text translation. The resulting transcript is used as the basis for natural language generation (Gonfalonieri 2018). The network used for natural language processing relies on machine learning techniques and is trained on annotated data. Amazon uses active-learning strategies to maximise the training efforts. The used network learns from failed enquiries in order to prevent future error. This self-learning technology is at the interface of AI, thus it can be said that the system is based on AI rather than being Artificial Intelligence itself (Horowitz 2020). Amazon has released the code to allow developers to program and add Alexa skills themselves ((Kim 2018)).

On this basis, one can conclude that the article's classification of Alexa is not far from the truth. Nonetheless, formulations such as "everyday AI" (Fisch 2020) are misleading and terms seem to be used without a general understanding of their meaning. This can be seen in the discussion about bias and Alexa (as outlined in Section 4.1.1), which does not accurately portray the different processes involved in language generation.

4.3.2 Emotive portrayal of AI voice assistant

To combine the investigation of the technical representation of Alexa with the discussion of tone as seen in Section 4.2, we decided to include some suggestive quantitative data on the correlation of "Classification of Alexa as AI" and "Portrayal of Alexa". An investigation of the occurrence of the subcodes of "Emotions" as well as "Portrayal of Alexa“ in general in the AI corpus and non-AI corpus did not reveal an evaluable difference in the overall distribution of codes between the two article corpora. Thus, one could conclude that there is no significant difference to be found in the overall portrayal of Alexa dependent on its classification as AI. Nevertheless, a closer look at the codes "Dangerous" and "Negative" reveals that there is a meaningful difference for some subcodes. Both the occurrence of these codes within articles and across articles of the AI-corpus is higher than for the non AI-labeled articles. Figure (1) shows the number of occurrences of the tags "Negative" and "Dangerous" for both corpora. This illustrates that the negative perception of Alexa correlates with Alexa being labeled as AI. Nevertheless, due to the scope of the collected data, only limited significance can be attributed to these quantitative findings.

Finally, a qualitative analysis of the tag "Dangerous" shows that the explicit depiction of situations as dangerous, while in general rare, is often part of quotes of agents with either professional or user experience. Some articles cite the UN report on gender bias (E.g. Dawson 2019, Rawlinson 2019, Petter 2019, Chowdhury 2019), while others cite lecturers (Cf. Lynskey 2019) or the co-founder of Apple (Cf. Fisch 2020). Examples of depiction of danger range from concerns about the smart speakers connection to bank accounts (BBC News 2020), over imposed gender biases (Dawson 2019: E.g.) to hacked microphones (Daily Mail 2019b). The general use of a voice assistant is however not seen as dangerous per se. One exception is a Guardian article (E.g. Lynskey 2019), which depicts the threat of a network of smart speakers across households, forming an infrastructure for surveillance. Nonetheless, perception of dangers is generally not found in the context of data security or privacy. Though the issue of data security is often mentioned, it is not inherently portrayed as dangerous. The only exception is the issue of bank account data, however never data in general.

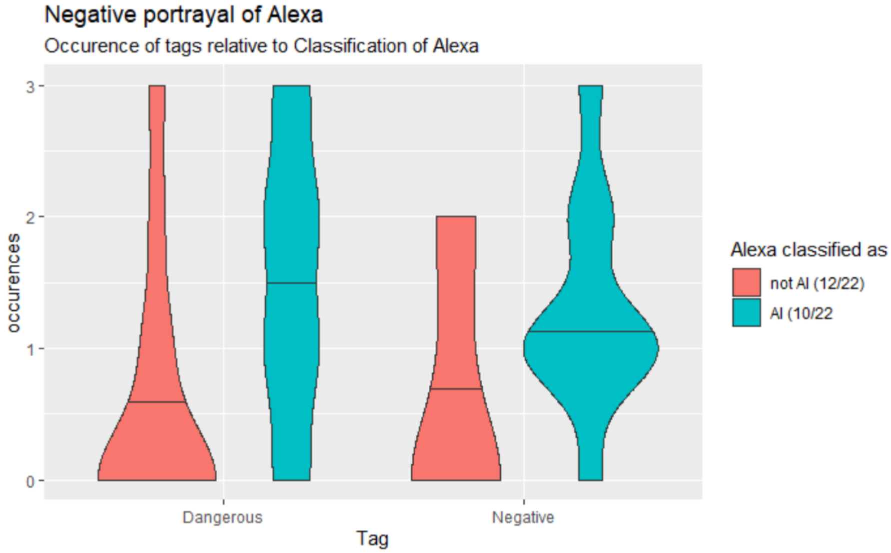


Figure 1: test

5 Discussion

On the basis of both the quantitative and qualitative analyses, further conclusions and evaluations of the discourse can be drawn. First of all, it is important to note that the articles do not refer to "ethics" explicitly (Cf. 4.1). In combination with the way that concerns are framed and the types of suggestions proposed, this suggests that the analysed news articles do not address overarching issues, but rather react to alerting revelations.

On the topic of privacy, the articles mainly express concern about personal data being reviewed by Amazon employees. The handling of data from technical devices is thus mainly recognized as a threat when human reviewers come into play. Similarly, concerns around transparency focus on the issue of human reviewers and mostly fall short to mention transparency issues with regards to how the data is used. This might be due to a limited understanding of the implementation and working of Alexa. Most articles do not set out to discuss technical details and they fall short of accurately describing how the voice assistant works. Nevertheless, the humanising vocabulary that is used to talk about "Alexa" as well as the use of "AI" as a description of Alexa without further comments paint a skewed picture. Finally, only half of the articles address the connection to AI, while the rest relies on labels such as “smart” and “intelligent”. The general understanding of the way that neural networks are trained and thus the importance of data for both development and further learning is limited. This might be why the discussion on data security mostly centers around “human reviewers” and disregards the general meaning of data for companies such as Amazon, especially in the context of developing AI. There seems to be little awareness about the implications of large scale data collection and the inherent value of data. This might be caused by a lack in background knowledge of journalists about AI.

In relation to bias, the articles focus on gender related issues. It is striking that the reason for gender-bias is mostly localised in the coding process and the high-pitched “female” voice of the assistants. While this is an important point to criticise, there seems to be very little understanding of the inherent biases present in the data on which networks are trained. This misunderstanding of bias could be traced back to the expectation that data and technology are neutral and finally, a lack of knowledge about machine learning. Instead, the problem is seen in humans who are portrayed to be failing. Amazon’s reviewers and other staff respectively become the reason for criticism. Finally, one reason why other forms of discrimination such as racial bias receive less attention might lie in the fact that racism is still often swept under the table on a societal level.

In terms of suggestions for data security concerns, the articles mainly focus on the ways in which users can control how their data is used. This is especially interesting as the proposed options are limited in their efficacy. The suggestion to “turn off Alexa” might do some good. However,it still falls far from being a reasonable guideline to protect data. This implies the underlying assumption that only some data needs to be protected, while in general there is no problem with data collection. The proposal to use the privacy settings to ensure data protection seems to fall short as the options provided have vague descriptions and do not allow for explicit control. The opt-out setting apparently allows users to stop the use of their data for further feature development (Cuthbertson 2019b), however this seems to only concern the learning processes involving human reviewers. Finally, users can only influence where and when they use Alexa and in a limited manner how the data collected in their household is used by Amazon. Thus, the trust is placed in Amazon’s handling of the user data as well as the development process in the case of bias, while the need for legal regulation of companies such as Amazon is rarely expressed. Thus, news articles do not advertise governmental action and legal regulation of the industry.

One main finding about the language use in the articles is the all dominating use of quotations. The articles mainly convey different opinions on the subject matter “Alexa” rather than explaining how ethical concerns come about. Thus, the portrayal of Alexa, emotions expressed and the attitudes presented are mixed. Quotations of former Amazon employees suggest expert expertise, however the actual statements are opinions by virtue of their user experience. These seem to be used nonetheless to convey expert advice on the proper use of Alexa. Additionally, the articles include a lot of statements by Amazon's spokesperson which are left without evaluative comments and thus, could either convince readers that Amazon handles the concerns well or be left with the feeling that Amazon's reaction is unsatisfactory. Finally, the overall attitude towards Alexa seems to be negative. However, all articles also propose suggestions either on how to continue using Alexa "safely" or on what to improve in the development of future voice assistants. Thus, there seems to be a general belief in the utility of systems like Alexa and an acceptance that they will be part of daily life moving forward. There are few suggestions to completely abandon the opportunities offered by self-improving assistants such as Alexa.

6 Conclusion

As outlined above, Alexa is discussed in media articles as an example of either a smart-speaker or AI-based voice assistant. Analysing the discourse by taking the example of Alexa has been illuminating. The investigations enabled us to identify how concerns about AI translate to technologies already in common use. The discussion of ethical concerns, while completely implicit, does cover the central questions at the heart of the Ethics of AI. Nonetheless, the focus of the UK news articles is directed by revelations concerning the practices of Amazon and investigations led by organisations. All in all, the discussion of concerns is focused on particular subfields and disregards several areas of ethical questions due to limited public understanding about machine learning. Issue that are not addressed include the biases inherent in training data, as well as the usage of collected data and the power gained by the company. One very central concern about the development of Alexa's features seems to be supervised learning workflows where data is analysed by human beings. The quantitative and qualitative analysis of the coded article corpus revealed recommendations for future development and use. These suggestions are mainly targeted at users and the company. While the general attitude towards Alexa is shaped by concerns and thus rather negative, for the most part the opportunities offered by assistant services are valued and the framing of the criticism opens up possibilities for future developments and adaptations. Finally, the discussion of ethical concerns in news articles offers important insights into concerns that consumers have about technologies based on AI and could be an influential factor in future product development.

7 Appendix

Table 3: Codeset

Code/Codeset	Description
Ethical questions:	Something is described as ethically concerning
Bias	Racial, sexist or classist biases (of AIs)
Data security	Surveillance, privacy or data security concerns
Human reviewers	Humans review of recordings of technical devices
Juridical guidelines	Juridical guidelines lead to an ethical problem
Technical mistakes	Technical mistakes lead to an ethical problem
Transparency	Procedures or policies are not transparent
Suggestions:	Proposed solutions for some ethical question
Develop ethical tool instead	Problems can be solved by ethical technologies
Juridical Complaints	Someone should file a suit
Legal prohibition	Something should be legally prohibited
User control	User should turn off, not use or carefully place Alexa
Responsibility of Amazon	Staff diversity, transparency, data protection/deletion
Portrayal of Alexa:	Alexa is portrayed in a certain way
Evaluation	Dangerous, negative, neutral or positive
Humanising	Alexa has gendered pronouns or listens, knows etc.
Amazon:	Critique or self-portrayal of Amazon
Language:	Style of the language used in the articles
Criticism of Alexa	Something about Alexa is critizised
Dramatization	Something is strongly exaggerated
Expert justification	Someone portrayed as an expert is cited
Uncertainty	Something is displayed as unclear or ambiguous
Emotions	Anger, doubt, fear, content or euphoria are expressed
AI:	Statements that relate to AI technology
Classification of Alexa as AI	Alexa is explicitly labeled as AI
Terms used	Smart speaker, machine learning, deep learning
Definition of AI	A definition of the term AI
Futuristic outlook	A prognosis of what will or could happen in the future

Table 4: Newspaper articles

Journal title	Title	Date
BBC News	Amazon Alexa: Luxembourg watchdog in discussions about recordings	2019-08-06
BBC News	Amazon Alexa security bug allowed access to voice history	2020-08-13
BBC News	Smart speaker recordings reviewed by humans	2019-04-11
The Guardian	’Alexa, are you invading my privacy?’ – the dark side of our voice assistants	2019-10-09
Guardian	Digital assistants like Siri and Alexa entrench gender biases, says UN	2019-05-22
Guardian	How to stop your smart home spying on you	2020-03-08
Daily Mail	Amazon’s Alexa and Apple’s Siri are SEXIST because its female voice reinforces the idea that women are ’subservient’, claims UN	2019-05-22
Daily Mail	WHY ARE PEOPLE CONCERNED OVER PRIVACY WITH AMAZON’S ALEXA DEVICES?	2019-04-11
Daily Mail	Is Siri sexist? UN cautions against biased voice assistants	2019-05-22
The Telegraph	I love my Alexa, but society will regret ignoring the dangers it raises	2020-02-25
The Telegraph	Smart speaker systems such as Siri and Alexa entrench gender bias, UN study finds	2019-05-21
The Telegraph	Amazon Echo customers could be in line for compensation over recordings, lawyer claims	2019-11-02
The Independent	Alexa should be banned from the bedroom, privacy expert says	2019-12-17
The Independent	Amazon admits employees listen to Alexa conversations	2019-04-11
The Independent	Amazon Alexa and Siri accused of sexism for ’thanking users for sexual harassment’	2019-05-22
The Mirror	How to stop Amazon employees ’listening in’ to your Alexa voice recordings	2019-08-06
The Mirror	Female voice assistants like Alexa promote idea that women are ’subservient’	2019-05-22
The Mirror	Amazon accused of ’spying on kids’ through Alexa-powered Echo speakers	2019-05-09
The Sun	SWITCHED OFF Ex-Amazon exec admits strangers DO listen to you and turns his Alexa off	2020-02-17
Daily Express	Artificial Intelligence: Expert warns of ‘potential disaster’ of AI bias	2020-09-20
Metro	Amazon launches compact Echo Show 5 with new focus on privacy	2019-05-29
Metro	Ex-Amazon exec admits he switches Alexa off for private conversations	2020-02-18

References

Agarwal, Anuj. 2021. *Top 70 UK news websites & influencers in 2021.* https://blog.feedspot.com/uk_news_websites/ (5 March, 2021).

BBC News. 2019. Smart speaker recordings reviewed by humans. *BBC News.* https://www.bbc.com/news/technology-47893082 (3 January, 2021).

BBC News. 2020. Amazon Alexa security bug allowed access to voice history. *BBC News.* https://www.bbc.com/news/technology-53770778 (3 January, 2021).

Chowdhury, Hasan. 2019. Smart speaker systems such as Siri and Alexa entrench gender bias, UN study finds. *The Telegraph.* https://www.telegraph.co.uk/technology/2019/05/21/female-voiced-smart-speaker-systems-entrench-gender-bias-un/ (9 January, 2021).

Curtis, Sophie. 2019a. Amazon accused of 'spying on kids' through Alexa-powered Echo speakers. *Daily Mirror.* https://www.mirror.co.uk/tech/amazon-accused-spying-kids-through-15020611 (10 January, 2021).

Curtis, Sophie. 2019b. How to stop Amazon employees 'listening in' to your Alexa voice recordings. *Daily Mirror.* https://www.mirror.co.uk/tech/how-stop-amazon-employees-listening-18838665 (10 January, 2021).

Cuthbertson, Anthony. 2019a. Alexa should be banned from the bedroom, privacy expert says. *The Independent.* https://www.independent.co.uk/life-style/gadgets-and-tech/news/alexa-privacy-amazon-echo-delete-recordings-a9249951.html (9 January, 2021).

Cuthbertson, Anthony. 2019b. Amazon admits employees listen to Alexa conversations. *The Independent.* https://www.independent.co.uk/life-style/gadgets-and-tech/news/amazon-alexa-echo-listening-spy-security-a8865056.html (9 January, 2021).

Daily Mail. 2019a. Is Siri sexist? UN cautions against biased voice assistants. *Daily Mail.* https://www.dailymail.co.uk/wires/ap/article-7060447/Is-Siri-sexist-UN-cautions-against-biased-voice-assistants.html (8 January, 2021).

Daily Mail. 2019b. Why are people concerned over privacy with Amazon's Alexa devices? *Daily Mail.* https://www.dailymail.co.uk/sciencetech/fb-6911113/WHY-PEOPLE-CONCERNED-PRIVACY-AMAZONS-ALEXA-DEVICES.html (8 January, 2021).

Dawson, Hannah. 2019. Amazon's Alexa and Apple's Siri are sexist because its female voice reinforces the idea that women are 'subservient', claims UN. *Daily Mail.* https://www.dailymail.co.uk/news/article-7056071/Amazons-Alexa-SEXIST-female-voice-reinforces-idea-women-subservient.html (8 January, 2021).

Fisch, Tom. 2020. Artificial intelligence: Expert warns of 'potential disaster' of AI bias. *Daily Express*. https://www.express.co.uk/news/science/1337896/artificial-intelligence-bias-warning-potential-disaster-of-ai-bias (12 January, 2021).

Flick, Uwe. 2013. *The SAGE Handbook of Qualitative Data Analysis*. SAGE Publications. 270. https://books.google.de/books?id=siIlCwAAQBAJ.

Gayle, Kesten. 2020. 15 mind-blowing stats about voice assistants. *Adobe Blog*. https://blog.adobe.com/en/publish/2020/09/21/mind-blowing-stats-voice-assistants.html#gs.uphz6p (15 January, 2021).

Gonfalonieri, Alexandre. 2018. How Amazon Alexa works? Your guide to natural language processing (AI). *Towards Data Science*. https://towardsdatascience.com/how-amazon-alexa-works-your-guide-to-natural-language-processing-ai-7506004709d3 (20 February, 2021).

Harris, Jamie & Shivali Best. 2019. Female voice assistants like Alexa promote idea that women are 'subservient'. *Daily Mirror*. https://www.mirror.co.uk/tech/female-voice-assistants-like-alexa-16181354 (10 January, 2021).

Horowitz, Jenn H. 2020. Is Alexa an AI? *IT chronicles*. https://itchronicles.com/artificial-intelligence/is-alexa-an-ai/ (5 February, 2021).

Johnston, Philip. 2020. I love my Alexa, but society will regret ignoring the dangers it raises. *The Telegraph*. https://www.telegraph.co.uk/politics/2020/02/25/love-alexa-society-will-regret-ignoring-dangers-raises/ (9 January, 2021).

Keach, Sean. 2020. Switched off – Ex-Amazon exec admits strangers do listen to you and turns his Alexa off. *The Sun*. https://www.thesun.co.uk/tech/10981326/amazon-alexa-spying-listening-bbc-panorama/ (3 January, 2021).

Kelion, Leo. 2019. Amazon Alexa: Luxembourg watchdog in discussions about recordings. *BBC News*. https://www.bbc.com/news/technology-49252503 (3 January, 2021).

Kim, Young-Bum. 2018. *The scalable neural architecture behind Alexa's ability to select skills*. https://www.amazon.science/blog/the-scalable-neural-architecture-behind-alexas-ability-to-select-skills (7 February, 2021).

Kinsella, Bret. 2019. Over 20% of UK households have smart speakers while Germany passes 10% and Ireland approaches that milestone. *Voicebot.ai*. https://voicebot.ai/2019/10/11/over-20-of-uk-households-have-smart-speakers-while-germany-passes-10-and-ireland-approaches-that-milestone/ (1 February, 2021).

Lynskey, Dorian. 2019. 'Alexa, are you invading my privacy?' – the dark side of our voice assistants. *The Guardian*. https://www.theguardian.com/technology/2019/oct/09/alexa-are-you-invading-my-privacy-the-dark-side-of-our-voice-assistants (5 January, 2021).

Müller, Vincent C. 2020. *Ethics of artificial intelligence and robotics*. Edward N. Zalta (ed.). Winter 2020. Metaphysics Research Lab, Stanford University.

Murphy, Margi. 2019. Amazon Echo customers could be in line for compensation over recordings, lawyer claims. *The Telegraph*. https://www.telegraph.co.uk/technology/2019/11/02/amazon-customers-could-receive-payout-illegal-recording/ (9 January, 2021).

Parsons, Jeff. 2019. Amazon launches compact Echo Show 5 with new focus on privacy. *Metro*. https://metro.co.uk/2019/05/29/amazon-launches-compact-echo-show-5-new-focus-privacy-9724048/ (11 January, 2021).

Parsons, Jeff. 2020. Ex-Amazon exec admits he switches Alexa off for private conversations. *Metro*. https://metro.co.uk/2020/02/18/ex-amazon-exec-admits-switches-alexa-off-private-conversations-12259105/ (11 January, 2021).

Petter, Olivia. 2019. Amazon Alexa and Siri accused of sexism for 'thanking users for sexual harassment'. *The Independent*. https://www.independent.co.uk/life-style/women/amazon-alexa-siri-sexist-un-women-ai-gender-stereotypes-a8924581.html (10 January, 2021).

Rawlinson, Kevin. 2019. Digital assistants like Siri and Alexa entrench gender biases, says UN. *The Guardian*. https://www.theguardian.com/technology/2019/may/22/digital-voice-assistants-siri-alexa-gender-biases-unesco-says (5 January, 2021).

West, Mark, Rebecca Kraut & Han Ei Chew. 2019. *I'd blush if I could: closing gender divides in digital skills through education*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000367416.

Winder, Davey. 2020. How to stop your smart home spying on you. *The Guardian*. https://www.theguardian.com/technology/2020/mar/08/how-to-stop-your-smart-home-spying-on-you-lightbulbs-doorbell-ring-google-assistant-alexa-privacy (5 January, 2021).

Chapter 15

How is AI explained to children? A qualitative analysis of educational videos for children

Franziska Gellert, Julia Laudon, Pia Münster & Rebekka Schlenker

Artificial intelligence is all around us, ranging from smart assistants to self driving cars or medical robots. In this rapidly developing field of science it has become increasingly important to teach society and especially children about what lies behind this technology. In this chapter, we conducted a qualitative analysis of different German educational TV shows in order to investigate how children are given an understanding of what AI is and how it works. The results can be taken as a short overview of how this is achieved with a special focus on application fields, style and tone of portrayal as well as possible future implications.

Keywords: Artificial Intelligence | Public Discourse | Documentaries | Children Documentaries | Education

1 Introduction

Nowadays, one can find applications of artificial intelligence (AI) almost everywhere. The technology is all around us, be it in our mobile phones or computers that beat even the world's best chess players at their own game (Gibbs 2017). The occurrences of AI in our world have risen dramatically in the last decades and will likely do so even more in the future (European Commission. Joint Research Centre. 2020). It has become increasingly important to teach society and especially children about AI, since they will often interact with it during their life. Therefore, they require explanations about what lies behind this technology, where they can find such implementations, what humanity can gain from it and also which problems and discussions might arise as a result. In the following we will discuss how children are given an understanding of artificial intelligence in the context of educational videos.

2 Methodology

To investigate how children are taught about AI, we conducted a manual, qualitative analysis of educational resources. We focused solely on German sources, as we felt like we could judge the quality and importance of those best. For this purpose, we found our material through the search engines YouTube, Ecosia and Google, via the keywords 'künstliche Intelligenz'/'KI' (artificial intelligence/AI), 'Kinder' (children)/ 'kindgerecht' (child-friendly) and 'Erklärung' (explanation)/ 'erklärt' (explained). We limited our material on educational videos as we found them to be quite exhaustive already and adding different media would have gone beyond the scope of what is possible in this work.

We excluded the videos that are published before 2018 in order to analyze only currently relevant resources. Furthermore, we excluded videos that do not explicitly mention AI and do not match the target group of children roughly between the age six and 14. Of the original 13 videos, we excluded five sources that did not match our inclusion criteria, leaving us with seven videos, all of which, except the video by bpb, are produced by German public broadcasting stations.

To capture our expectations and findings, we created a questionnaire with different subtopics that we answered while watching the documentary (1). We split our group into two subgroups and divided the videos equally. We then watched the videos repeatedly to answer the questionnaire and include timestamps. After this, we exchanged our findings and thereupon began our writing process.

3 Short description of Videos

3.1 Checker Tobi

'Checker Tobi' is an educational children's show and produced by to the public broadcasting station ZDF (Zweites Deutsches Fernsehen). Each video is about 25 minutes long and focuses on the explanation of one specific topic. The program aims at portraying and answering three questions about the subject. The

Table 1: Questionnaire

Topic	Questions
Depicted areas	Which ones are depicted? Is it done in a more child-friendly manner?
Level of detail	What is the difference to normal computers? Details: Are neural networks/the imitation of human brain mentioned?
Strengths and weaknesses	Is strong vs. weak AI explained? Is AI presented as perfect or faulty? Is task-specificity/efficiency brought up?
Humanization of AI	Are human verbs or adjectives attributed to AI? In how far differs AI from humans? (especially regarding emotions) How is the relationship between AI and humans depicted? Is a superiority/inferiority recognizable?
Future impacts	Which positive and negative future impacts are mentioned? To what extend are emerging ethical discussions mentioned?

video, “The artificial intelligence check” (“Der künstliche Intelligenz-Check” *Der Künstliche Intelligenz-Check* 2020), raised questions such as: ”How does AI work? How does an AI recognize emotions? What does AI look like in the future?” (*Der Künstliche Intelligenz-Check* 2020 2:50) To answer these questions, Tobi visits different people and places, two students at the Technical University Munich, an expert at the house of innovation in Munich and a futurologist, who inform him about different aspects of the topic.

3.2 Erde an Zukunft

’Erde an Zukunft’ is a children’s program by KIKA, a channel of the public broadcasting stations ARD (Arbeitsgemeinschaft der öffentlich-rechtlichen Rundfunkanstalten der Bundesrepublik Deutschland) and ZDF targeted at children. The show explores topics that will play a role in the future and explainig these topics to children. In the eleven minute long episode ”Artifical intelligence” (”Künstliche Intelligenz” *Künstliche Intelligenz* 2021), the reporter explains AI and tries to imagine a possible future where AI has been vitally integrated into human life.

3.3 KABU

’KABU’ is an explanation series by the Federal Agency for Civic Education, bpb (Bundeszentrale für politische Bildung), that intends to give children an understanding of difficult concepts. The four minutes long video ”Artificial Intelligence - explained for children” (”Künstliche Intelligenz - kindgerecht erklärt” Studio im Netz München 2020) introduces the topic of artificial intelligence by explaining fundamental facts and is published online on the website of bpb.

3.4 logo!

’logo!’ is a daily news program for children that explains different topics in short video clips and belongs to ZDF. The video clip on artificial intelligence, “What is “artificial intelligence”?” (“Was ist künstliche Intelligenz”? *Was ist ”Künstliche Intelligenz”? - logo! erklärt* 2018) is one minute and thirty seconds long and focuses only on the basics of that topic.

3.5 Löwenzahn

’Löwenzahn’ is an entertaining and educational children’s show and belongs to ZDF. Every video is about 25 minutes long and tells stories about the life of the main character, Fritz Fuchs, who is eager to learn about different topics. The episode about artificial intelligence, “Intelligence on wheels – the kidnapped Schlauto” (“Intelligenz auf Rädern – Das entführte Schlauto” *Löwenzahn: Intelligenz auf Rädern* 2021), shows Fritz Fuchs and his friend inventing and developing a self-driving robot, named Schlauto, that delivers pizza.

3.6 Neuneinhalb

’Neuneinhalb’ is a news program for children by the public broadcasting station WDR (Westdeutscher Rundfunk), which is a subsidiary of ARD. It addresses current topics in society in a way that is suitable for children. In the nine minute long episode ”Artificial Intelligence - How smart are machines?” (”Künstliche Intelligenz - Wie schlau sind Maschinen?” *Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019), the reporter explains what AI is by talking to different experts and asking children about their opinion on it.

3.7 PUR+

'PUR+' is a discovery magazine and belongs to ZDF. Each 25 minute episode shows the reporter Eric Mayer exploring topics from all over the world with a special focus on including children's ideas and opinions. In the episode "Can computers think?" ("Können Computer denken?" *Können Computer denken?* 2020), Eric gets to know different forms of AI and executes his own AI experiment with a school class.

4 Findings

4.1 Depicted Areas and Level of Detail

With the first part of our questionnaire, we want to take a closer look at which application areas of AI would be depicted in children's documentaries and how exhaustive this portrayal would be. One can assume that the choice of topics will be appropriate for children and therefore, certain areas will most likely have to be excluded, for example AI in the military force. We found that almost all of our researched material depicted AI in cars as a prime example of instantiated AI. Not only is AI in cars explained with the example of self-driving cars (*Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 6:16; Studio im Netz München 2020 3:12; *Künstliche Intelligenz* 2021 1:22; *Der Künstliche Intelligenz-Check* 2020 1:49, 4:16; *Löwenzahn: Intelligenz auf Rädern* 2021 5:22, 7:17) but also with AI in a lane keeping assist (*Löwenzahn: Intelligenz auf Rädern* 2021 7:29) , as automatic breaks (*Löwenzahn: Intelligenz auf Rädern* 2021 7:41) and in a parking assistance (*Löwenzahn: Intelligenz auf Rädern* 2021 8:00).

Next to this application field, we find a frequent reference to AI in smartphones, more specifically personal assistance systems like Siri, Google or Alexa (*Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 7:16; Studio im Netz München 2020 2:23; *Können Computer denken?* 2020 10:14; *Der Künstliche Intelligenz-Check* 2020 0:41, 2:08), as well as speech recognition and voice control (*Löwenzahn: Intelligenz auf Rädern* 2021 10:50). Furthermore, we observe a frequent mentioning of robots. Certainly, not every existing robot is controlled by an AI, but still, it is the application that most likely comes first to mind for most children.

A wide range of robots is mentioned throughout all of our video material, from the cooking robot PR2 in PUR+ (*Können Computer denken?* 2020 0:45) through the humanoid Pepper (neuneinhalb 0:36) to Sophia the robot (*Können Computer denken?* 2020 9:41). Besides these indisputable robotic instantiations, smaller robots, like vacuum cleaners and lawnmowers (*Löwenzahn: Intelligenz auf Rädern* 2021

3:42) and robotic assistance in medical care, surgeries, and prosthetics (*Der Künstliche Intelligenz-Check* 2020 8:49, 14:40; *Künstliche Intelligenz* 2021 1:30; *Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 0:49) are listed. Other applications that are mentioned, in order of their occurrence frequency, are:

- industry (*Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 6:58; *Löwenzahn: Intelligenz auf Rädern* 2021 4:04),
- chess or GO computers (Studio im Netz München 2020 0:57; *Können Computer denken?* 2020 9:00; *Künstliche Intelligenz* 2021 0:27),
- smart home (*Der Künstliche Intelligenz-Check* 2020 16:00; *Künstliche Intelligenz* 2021 1:25),
- face and/or emotion detection (*Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 4:49; *Der Künstliche Intelligenz-Check* 2020 9:43),
- intelligent assistance for astronauts (*Können Computer denken?* 2020 9:15) and
- chatbots (*Können Computer denken?* 2020 11:26).

With all this information on hand and with respect to the previously stated question about the exhaustiveness of the depiction, we want to briefly discuss how broad the field of application is represented in our assorted documentaries.

First of all, we are aware that a complete depiction of all areas of application would be out of scope for a short documentary, as well as inappropriate for younger children. Across all of the videos, the presented picture of AI applications is quite broad. Application of AI that are generally more present in a child's life, e.g. cars and smartphones, are mentioned in most of our material. In contrast, less immediate instantiations are only mentioned briefly, e.g. chess computer. Furthermore, some videos showed a deeper exhaustiveness than others, but that might be due to the different aims and time-limits of the shows.

Moving on, we will take a look at how AI is explained to the viewer. We can expect that the explanation will lack some details to make it more appropriate for the target group of elementary school children. Foremost, AI is explained as being a computer program that is written by humans: "Artificial intelligence is a system of computers, that can be really big or really small, like in our smartphone" ("Künstliche Intelligenz ist ein System aus Computern. Die können riesig groß oder mini-klein sein, wie in unserem Handy.“ *Was ist ”Künstliche Intelligenz”? - logo! erklärt* 2018 0:13).

We now turn to the question of what exactly makes an AI so fundamentally different from a ”common” computer program and how this distinction is explained. Our researched material congruently explains this distinction by the AI being able to learn. The documentary PUR+ explains this as such: ”Machines can learn to act independently, if we feed them with our knowledge.” (“Maschinen können also lernen, eigenständig zu handeln, wenn wir Menschen sie mit unserem Wissen füttern.“ *Können Computer denken?* 2020 8:05). Therefore, the important distinction from an AI to a “common” computer program is that intelligent programs can learn independently (*Können Computer denken?* 2020 8:43; *Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 2:14, 2:05, 4:16; *Künstliche Intelligenz* 2021 0:53; *Der Künstliche Intelligenz-Check* 2020 3:49, 7:55; Studio im Netz München 2020 1:36). Besides this learning ability, intelligent programs should be able to solve problems (*Was ist ”Künstliche Intelligenz”? - logo! erklärt* 2018 0:47) and learn from their own mistakes (*Was ist ”Künstliche Intelligenz”? - logo! erklärt* 2018 1:00).

It is frequently stated that an AI does not need to be one sole computer that does all the work, but instead a system of several intelligent computers that work together (*Was ist ”Künstliche Intelligenz”? - logo! erklärt* 2018 0:12, 1:11; *Der Künstliche Intelligenz-Check* 2020 1:59, 4:11). Again, self-driving cars are mentioned as an example of such cooperative work between several AI’s (*Der Künstliche Intelligenz-Check* 2020 4:11).

Now that the distinction of AI from typical programs is clarified, we turn to the question of what exactly a “perfect” AI aims to be. In the case of chatbots in PUR+, it is stated that „if a system acts as if the answer could be from a human, then one could denote it as artificial intelligence” (“wenn sich das System tatsächlich so verhält, dass die Antwort von einem Mensch kommen könnte, dann kann man es als künstliche Intelligenz bezeichnen“, *Können Computer denken?* 2020 18:18). Here, the viewer is told that an indiscernible imitation of human behavior equals so-called “intelligence”. This imitation of the human brain as the ultimate goal for an AI is either explicitly (*Was ist ”Künstliche Intelligenz”? - logo! erklärt* 2018 0:21; *Können Computer denken?* 2020 8:24) or implicitly (*Löwenzahn: Intelligenz auf Rädern* 2021 7:23, 10:44; Studio im Netz München 2020 1:09) mentioned in most of our researched material. However, it is also frequently emphasized that this goal is not fully reached yet.

Regarding the level of detail that is displayed in the explanation of AI, we find that most of our material agrees in the rather superficial level of explanation. Still, some of the videos go into a bit more detail as to how such a program is implemented (*Können Computer denken?* 2020 11:26) or what the training process of an AI can look like (*Können Computer denken?* 2020 5:55). Even educational videos for children need to have at least some sort of entertainment factor to keep the young viewers' attention. Thus, we can expect to observe a trade-off between the level of detail in the explanation and the entertainment aspect in the given time limit of the respective shows. If we take all of our material together, the explanation of AI is quite well-rounded. All the necessary and important details are explained on a level that seems appropriate for the age group. For example, we notice that a common way of hinting at neural networks as an underlying type of algorithm is by comparing the AI to the human brain and noting exactly such as the goal of a "perfect" AI. However, we notice that some of the videos are missing significant aspects of AI in their explanation. Take for example the video "Löwenzahn". Here, the technical explanation of the machine in which the AI is implemented stands out more than the explanation of the AI itself (e.g. *Löwenzahn: Intelligenz auf Rädern* 2021 6:22, 10:10, 12:14). As mentioned earlier, this video is more focused on the entertainment aspect and thus the more detailed explanation of AI falls shorter than in other videos.

4.2 Strengths and Weaknesses

Another topic of interest to us is how the different videos address and portray the strengths and weaknesses of artificial intelligence. We asked ourselves if and how the notion of a strong versus a weak AI is discussed, whether artificial intelligence in general is portrayed as superior or inferior to humans and consequently, how this might steer the overall perception of AI. The dominant style of explanation in all videos is the contrasting comparison of a 'perfect', omnipotent machine that is capable of doing everything compared to machines that are very task-specific and valuable only in their designated field of application but lack a lot of skill in other areas.

Simple examples of the superiority of AI that are frequently mentioned throughout all of the videos include that of computational power or the different gaming AIs such as Alpha GO: "When it comes to storing and processing huge amounts of data, computer programs are already even better than humans." ("Wenn es darum geht, riesige Mengen an Wissen zu speichern und zu verarbeiten, sind Computerprogramme sogar schon besser als wir Menschen." *Können Computer denken?* 2020 8:35). An essential advantage which is often mentioned is that, with the help of AI, it becomes possible and more accessible to do work that is too dangerous for humans to do, such as doing research in deep sea or in space (e.g. Studio im Netz München 2020 2:43). AI is hence perceived as an asset or 'helper'. This example is an indicator for the different types of relationships that are displayed between humans and AI which we will come back to later on (see Humanization of AI).

Furthermore, the majority of the videos focus on one specific instance of AI which is oftentimes some type of robot. The AI is shown while performing a particular task that it has mastered. In PUR+ for example, the robot PR2 is observed while doing a sequence of actions that lead to it being able to cook popcorn independently (*Können Computer denken?* 2020 1:16). In that way, the viewer gets the impression of an autonomous machine that can successfully work on its own, though not as smoothly and naturally as humans.

On the other hand, the same application of AI is shown performing more difficult tasks, or rather those that it has not yet learned. In these situations, it is obvious that the AI in question is not perfect and can even fail at the simplest task. Having a look at the example in PUR+ again, they show PR2 stuck in the process of assembling things for eating muesli with the spoon hidden in a drawer (*Können Computer denken?* 2020 3:02).

Another fitting example for this type of opposition can be found in the same video PUR+. In cooperation with a secondary school, they perform an experiment with a chatbot AI. This is done in style of a Turig-test, i.e. the children formulate questions which are either answered by the chatbot AI or the moderator. Afterwards they have to label the answers accordingly. Interestingly, although the students are quite confident to be able to distinguish between chatbot and human, in the end, the chatbot is able to deceive them 14 out of 54 times which they rate to be a surprisingly good quota (*Können Computer denken?* 2020 18:50).

Other examples of inferiority include smart assistants that are not able to comprehend personal questions or robots that fail at face recognition (*Können Computer denken?* 2020 10:48; *Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 5:30). This error-proneness is often made clear implicitly by repeatedly showing how AIs are faulty. In the video Checker Tobi for example, the central AI which they called “Waltraud” never really masters the ability of properly talking to the moderator or fulfilling the given tasks correctly (e.g. *Der Künstliche Intelligenz-Check* 2020 1:08). In this context, it is also often mentioned that it is desirable to constantly optimize AI (e.g. *Was ist ”Künstliche Intelligenz”? - logo! erklärt* 2018 1:22) in order to make it smarter, even more intelligent and possibly someday even as smart as humans. This is another aspect we will discuss in more detail in a later section (see Future Impacts).

Accordingly, most videos also allude to the idea that there might someday be some kind of artificial intelligence that will be just as smart as humans or maybe even smarter: “Who knows? Maybe they will someday be just as smart as us” (“Wer weiß? Vielleicht werden sie irgendwann wirklich so schlau sein wie wir Menschen" *Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 9:04). It is interesting to note that only one of the videos explicitly mentions the difference between strong and weak AI by name. As seen above, the differentiation is often made implicitly, but the terms "strong" and "weak" AI are not specifically discussed (except for the video KABU). We perceived this with surprise, as it is arguably not a very difficult definition/concept for children to understand. One could even suggest that actually denoting the difference as "strong" versus "weak" could possibly enhance the comprehension of it.

Apart from that, all videos put a special focus on clarifying that this kind of perfect, intelligent machine does not exist yet and that there is still a long way to go before this will be the case: "So eine perfekte Form von künstlicher Intelligenz gibt es noch nicht" (*Was ist "Künstliche Intelligenz"? - logo! erklärt* 2018 1:22). This point is often illustrated by emphasizing how AI, as of right now, differs from humans, namely that humans are considerably more flexible in everything that they are able to do (*Können Computer denken?* 2020 19:50). This implies that humans have a lot of different skills, as opposed to just being good at one particular task, and, importantly, that they can do them at the same time (*Können Computer denken?* 2020 20:20). These abilities are pointed out as easy and very intuitive for humans but almost impossible for machines, such as coordinating opening a door with a key (*Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 2:32). Additionally, AI is said to lack emotions and emotional intelligence (e.g. *Der Künstliche Intelligenz-Check* 2020 13:15) which is defined as "the ability to understand the way people feel and react and to use this skill to make good judgements and to avoid or solve problems" (Press 2014). This is another factor distinguishing them from human beings as we will again see in more detail later on (4.3).

Overall, it can be concluded that almost all of the videos try to implicitly show the contrast of a strong versus a weak AI. By comparing different levels of capability of AIs, it is made clear where AI, as of right now, is superior or inferior to humans and why. The overall image that is thus created is that AIs, currently, are becoming more and more advanced but also still hold a lot of weaknesses that leave room for future research and ever-growing optimization.

4.3 Humanization of AI

Furthermore, we are particularly interested in what way the documentaries present AI in order to make the topic more appealing to children. Most content for children shows a clear tendency of humanizing animals and inanimate objects

You 2020. Therefore, the question arises whether artificial intelligence would be especially humanized in the documentaries for children.

To answer this question, we first investigate the words that are attributed to AI. In most videos, verbs that are usually attributed to humans are ascribed to instances of artificial intelligence. Verbs like:

- "to figure something out" ("kapieren", *Löwenzahn: Intelligenz auf Rädern* 2021 7:25),
- "to keep in mind " ("im Kopf behalten", *Löwenzahn: Intelligenz auf Rädern* 2021 08:55),
- "get an idea of it" ("macht sich ein [...] Bild von",*Löwenzahn: Intelligenz auf Rädern* 2021 7:01),
- "he knows" („weiß er" *Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 3:50),
- "decides" („entscheidet sich" *Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 3:46) and
- "understanding" („verstehen" *Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 7:07)

give the impression that artificial intelligence thinks just like humans do. In the video Checker Tobi the verbs "care about" ("kümmert sich",*Der Künstliche Intelligenz-Check* 2020 04:20), and "argue" ("streiten",*Der Künstliche Intelligenz-Check* 2020 20:05) are ascribed to artificial intelligence, leading the viewer to believe that there is an emotional reasoning behind the AI's actions. The same can be seen in the video Löwenzahn, where the main character rides a self-driving bus and the AI asks him several times to sit down, until it seems to get impatient: "please finally sit down" ("bitte setzen Sie sich endlich hin",*Löwenzahn: Intelligenz auf Rädern* 2021 5:38).

In some of the analyzed videos, also adjectives that are normally used to describe humans are ascribed to instances of AI. Examples are:

- „decent" ("anständig", *Können Computer denken?* 20205:42),
- „self-reliant" ("selbstständig", *Löwenzahn: Intelligenz auf Rädern* 2021 8:15) a nd
- „smart" ("schlau", *Löwenzahn: Intelligenz auf Rädern* 2021 7:52; *Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 0:27).

In some videos, humans gave the artificial intelligence names or nicknames, like “Waltraud” (*Der Künstliche Intelligenz-Check* 2020 0:52), “Sweetie” (“Süßer”, *Löwenzahn: Intelligenz auf Rädern* 2021 10:49) or “Baby” (“Baby”, *Löwenzahn: Intelligenz auf Rädern* 2021 11:04). In general, a lot of the used wording suggests that artificial intelligence interacts and, as already stated, thinks just like humans. This can be seen for instance in the video Löwenzahn:

> “It [the AI] has noticed the tired and jerky movements of the driver and reports: ’Hello, time for a coffee break’”
>
> (“Es hat die müden und ruckartigen Bewegungen des Fahrers gemerkt und meldet: ‘Hallo, Zeit für eine Kaffeepause‘“, *Löwenzahn: Intelligenz auf Rädern* 2021 7:53)

or in a scene in Checker Tobi, where the interaction between different AIs of a Smart Home is described as following:

> ”The refrigerator recognizes: ‘the milk is empty’, the coffee machine sends: ‘I’m ready!’, and the heater says: ‘Everyone is out of the house? I’ll turn down the heat, or is that too cold for you, toaster?’”
>
> („Der Kühlschrank erkennt: „die Milch ist leer“, die Kaffemaschine sendet: “bin bereit!“, und die Heizung sagt: „Alle sind aus dem Haus? Ich stell mal die Heizung runter, oder ist dir das zu kalt, Toaster?“,*Der Künstliche Intelligenz-Check* 2020 17:41).

Wordings like this give the impression that robots are more human-like than they actually are.

Some videos, however, emphasize the clear distinction between an AI and a human being. For instance: “sure, artificial means not human” (“klar, künstlich heißt nicht menschlich”,*Was ist ”Künstliche Intelligenz”? - logo! erklärt* 2018 0:10) and ”It [the AI] is just not a human” (“Das [die KI] ist einfach kein Mensch“,*Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 06:02). Nevertheless, the abilities of artificial intelligence are often compared to those of humans. One video points out that AI learns from mistakes and feedback, just like humans (*Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 04:00). It is also mentioned that humanoid robots, like Pepper, resemble the appearance of humans (*Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019, 0:36). In most videos it is brought up that an important ability that differentiates AI from humans is having emotions (e.g. Studio im Netz München 2020 1:21; *Können Computer denken?* 2020 12:55;

Der Künstliche Intelligenz-Check 202013:15; *Löwenzahn: Intelligenz auf Rädern* 2021 18:51).

Another interesting aspect we consider in our analysis is the way the relationship between humans and AI is portrayed. Most videos show AI as a helper or servant for humans. AI can help in the household (*Der Künstliche Intelligenz-Check* 2020 18:04), with transportation as in self-driving cars (*Der Künstliche Intelligenz-Check* 2020 07:46; *Löwenzahn: Intelligenz auf Rädern* 2021 7:23), in the health sector as care-givers or assistants to doctors (*Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 0:50; *Der Künstliche Intelligenz-Check* 2020 14:43), or to support humans in doing their work (*Können Computer denken?* 2020 0:26, 9:25; *Der Künstliche Intelligenz-Check* 2020 22:31). For instance, the video Löwenzahn revolves around the story that the main character builds a self-driving delivery robot that helps delivering pizza for the local pizza shop and can thereby save the pizza shop from going bankrupt (*Löwenzahn: Intelligenz auf Rädern* 2021).

AI is not only portrayed as a major help for humans but also as a friend or companion. Especially speech assistants and chatbots are described as virtual friends for humans (*Künstliche Intelligenz* 2021 3:24; *Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 7:21), though it is mentioned that they cannot replace real friends (*Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 8:58). In PUR+ for example, AIs that play with humans are depicted (*Können Computer denken?* 2020 0:23). In Checker Tobi, there is a sequence where a robot tells Tobi, the main character who is clearly over-worked, to take a break (*Der Künstliche Intelligenz-Check* 2020 13:35) and brings him a tea when he is sick (*Der Künstliche Intelligenz-Check* 2020 13:45). Though this scene is only supposed to display how an interaction between humans and robots could look like in the future, it clearly shows an empathetic and personal interaction that could resemble a friendship or even a parent-child relationship, where the robot takes care of the human. A more extreme and reversed version of the parent-child relationship between AI and human is conveyed in the video Löwenzahn. Here, the two main characters call the self-driving delivery robot they invented their "baby" ("Baby", *Löwenzahn: Intelligenz auf Rädern* 2021 13:46) and refer to it as "Bambino" ("Bambino", *Löwenzahn: Intelligenz auf Rädern* 2021 15:37), and to themselves as "mum and dad" ("Mama und Papa", *Löwenzahn: Intelligenz auf Rädern* 2021 23:32). The notion of a human as the creator of an AI, for instance mentioned in the video KABU (Studio im Netz München 2020 0:40, 1:44), suggests a hierarchical relationship, in which the human is superior to the AI. None of the videos portrays AI as fully superior to the human, at most as superior in a certain task, like playing chess (Studio im Netz München 2020 1:03).

To sum up, most videos make, in principle, clear that AIs are distinct from humans. However, the AIs in the videos are often humanized by attributing words with certain notations to them, giving them human names and portraying a relationship with humans that resembles typical relationships among mankind, like a friendship or parent-child relationship.

This humanization is, however, not unexpected, as it is included in most content for children since it makes topics more appealing, interesting and understandable. The notion of AI as a friend or helper for humans, which is present in most of the analyzed videos, conveys a positive attitude towards AI and makes the topic more accessible. However, the, in some videos very extreme, representation of AI as human-like can be seen as problematic, as it can immensely influence the picture children have of AI, which however does not correspond to reality. Therefore, it could lead to children making incorrect assumptions about AI. The documentaries have to find a good balance between the accuracy of the information and the entertainment of the children in order to avoid conveying an unrealistic picture of AI. For some of the analyzed video, for instance Löwenzahn, the focus lays more on the entertainment aspect while others, like logo!, focus more on the education aspect.

4.4 Future Impacts

Finally, we have a closer look at how and to what extent potential future impacts are mentioned in the documentaries. Firstly, we can notice that not all videos address the future impacts that AI could potentially have on society and especially the ethical concerns that come along with it. However, some of the documentaries, especially the video Checker Tobi, deals with positive and negative future impacts as well as emerging ethical discussions and those videos will be analyzed in the following.

Throughout the videos it is often emphasized that AI has become an integral part of our lives and, due to its rapid development, will become ever more important in our future lives (Studio im Netz München 2020 00:20, 3:34; *Was ist "Künstliche Intelligenz"? - logo! erklärt* 2018 1:23). A very prominent area mentioned in some documentaries is that AI will take over an ever-increasing number of tasks in the future that were originally carried out by humans (*Der Künstliche Intelligenz-Check* 2020 8:36, 21:52, 22:32). For instance, it is stated that AI will further be utilized in the medical sector, where it is already used to improve prostheses (*Der Künstliche Intelligenz-Check* 2020 14:54) or to support doctors in detecting certain diseases which can lead to a faster recovery (*Der Künstliche Intelligenz-Check* 2020 14:42; *Künstliche Intelligenz* 2021 5:37). It is also mentioned that, if AI would be capable of detecting emotions adequately, it "could also recognize whenever someone does not feel good and then take care of them" ("[...] würde auch erkennen, wenn es uns nicht gut geht und sich um uns kümmern" *Der Künstliche Intelligenz-Check* 2020 13:40).

Although this often implies a relief for humans, many people are concerned that they could lose their job. This issue is broached in several of the documentaries which underlines the prominence of this fear in society (*Der Künstliche Intelligenz-Check* 2020 8:52; *Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 5:51; *Künstliche Intelligenz* 2021 5:48). For instance, in the video neuneinhalb this worry about job-loss is incorporated into a conversation as the reporter ironically expresses her fear of losing her job as a response to the robot "Pepper" introducing himself as the new reporter (*Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 0:16). According to the documentary Checker Tobi, this is a realistic concern but only for tasks that can be easily carried out by an AI (*Der Künstliche Intelligenz-Check* 2020 8:59) including very repetitive work like delivering parcels or building cars (*Der Künstliche Intelligenz-Check* 2020 8:35). Moreover, it is emphasized that AI can bring new job opportunities due to additional tasks that are emerging (*Der Künstliche Intelligenz-Check* 2020 9:02). Consequently, although the question is often not explicitly answered, AI can rather be seen as an addition to instead of a replacement of humans (*Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 5:50; *Können Computer denken?* 2020 19:30) and could give people time that they can use more efficiently (*Der Künstliche Intelligenz-Check* 2020 22:00).

As observed before, self-driving cars are frequently mentioned throughout the material and often used as a prime example for emerging ethical discussions in the context of the future use of AI. One critically discussed concern in this debate is whether it could be too dangerous to let self-driving cars decide over human lives on their own (*Der Künstliche Intelligenz-Check* 2020 7:26). In the documentary neuneinhalb, they dodge this question by declaring that "there is just such a susceptibility to errors. One has to program very carefully" ("Es gibt einfach so eine gewisse Fehleranfälligkeit. Da muss man halt sehr vorsichtig programmieren" *Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 6:36). Generally, even though some videos indicate that there need to be more laws and regulations in order to use self-driving cars, or AI in general, in real life (Studio im Netz München 2020 3:25; *Der Künstliche Intelligenz-Check* 2020 20:17), this aspect of artificial intelligence is not discussed in much detail.

Concerning the relationship between humans and AI, it is often stressed that AI should never decide completely on its own and final decisions should always be made by humans (*Künstliche Intelligenz* 2021 9:46; Studio im Netz München

2020 3:42). One possible way of how this could be realized is depicted in Erde an Zukunft, where they refer to an emergency off button that scientists in Oxford are currently working on (*Künstliche Intelligenz* 2021 10:00). In line with this, the video Checker Tobi introduces Isaac Asimov's rules for cohabitation between AI and humans to ensure that humans are the ones staying in control (*Der Künstliche Intelligenz-Check* 2020 20:30).

As we have already seen in the subsection Strengths and Weaknesses, some documentaries hint to the idea that AI could someday potentially become just as efficient and perhaps even smarter than humans (*Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 9:01; *Können Computer denken?* 2020 20:22). Linked to this is the issue of whether one should be afraid of that. If this question is addressed at all in a video, it is in a way that leaves the decision with the viewer, e.g. in the video PUR+: "What do you think, should we be afraid of it or is it just super cool?" (”Muss uns das Angst machen oder ist das einfach nur super cool? Was glaubt ihr denn?" *Können Computer denken?* 2020 20:30) and neuneinhalb "What do you think? Is it great or does it make you afraid?" (”Was denkt ihr? Ist das was Gutes oder macht euch das eher Angst?“ *Künstliche Intelligenz – Wie schlau sind Maschinen?* 2019 9:06). Moreover, even though the future development of AI is generally depicted in a rather positive way, in one video they also use more negatively connotated words like "quite scary" ("schon ziemlich unheimlich" *Künstliche Intelligenz* 2021 02:01), which could perhaps implicitly give children the impression that they actually should be afraid of it.

Two of the documentaries also try to give the viewer an impression of how the future with AI could look like. The documentary Erde an Zukunft depicts a simulation where they try to show a future in which AI is fully incorporated into our daily lives. In the video an AI controls the whole day of the human by, for instance, telling him what clothes to wear (*Künstliche Intelligenz* 2021 7:25) and managing his leisure time (*Künstliche Intelligenz* 2021 8:37). This is, in our opinion, a very exaggerated, stereotypical and unrealistic depiction that rather resembles a science fiction scenario. Moreover, even though the AI is displayed as very smart and powerful, there are also instances in which it becomes clear that even in this potential future, AI is not perfect and that it is not desirable that it should control our lives, which becomes clear in a scenario where the AI takes a figure of speech in a literal sense and breaks the protagonist's furniture: ”You said you wanted to really enjoy yourself again" (”Du hast gesagt, du wolltest es mal wieder richtig krachen lassen" literal translation: ”You said you really wanted to let it crash again", *Künstliche Intelligenz* 2021 9:24). Another situation that shows the future in a rather stereotypical manner can be seen in Checker Tobi where the reporters are 'beamed' 20 years into the future and suddenly wear silver, or metallic, clothes (*Der Künstliche Intelligenz-Check* 2020 19:00). However, this seems more like a humorous depiction of the future since the moderator asks "Do you really think that people will look like this in the future?" („Meinst du so sehen die Leute in der Zukuft aus?" *Der Künstliche Intelligenz-Check* 2020 19:14) and afterwards, they switch back to their old clothes.

To summarize, it can be noted that, despite the fact that it is always expressed that AI will play a huge role in the children's future, surprisingly few videos broach the topic of possible future impacts of AI on society, especially negative ones. Emerging ethical issues, such as those related to autonomous driving, are also addressed in only a minority of the videos. However, it is of course debatable whether such aspects should be dealt with exhaustively in documentaries that are specifically made for children. A few videos, however, try to address future implications, mainly by talking about job loss or self-driving cars. Additionally, some videos aim at encouraging the viewer to further think about the topic on their own which can be great way to get children involved in this discussion.

5 Conclusion

Bringing together the different aspects that we inspected in our analysis, we can summarize our findings as follows.

Firstly, it can be noted that the videos do not differ too much in their explanation of AI and were generally quite informative and well produced. Unsurprisingly, most videos make use of very popular and child-friendly applications of AI such as self-driving cars and various forms of robots. However, they also explain that AI can be found in simpler devices as well, for example smartphones. Considering the age group that these videos are targeted at, we find the level of detail, for example how AI learns, quite appropriate and well portrayed.

Moreover, all videos make an, at least implicit, distinction between strong and weak AI and try to illustrate the difference through many examples. In this context, it is often mentioned that a strong AI does not exist as of now but could potentially someday in the future. Another point that the majority of videos have in common is the fact that AI is heavily humanized, especially in the case of robots. At the same time, this often conveys AI as a 'friend' or 'helper' for humans. Concerning future implications and potential discussions, we are surprised at how little this is spoken about, especially considering that it will undoubtedly play a huge role in the future of today's children. All in all, we find the educational mission coupled with the entertainment aspect of the videos to be very well balanced.

6 Limitations and Outlook

We want to conclude our chapter by pointing out the limitations of our analysis as well as to give an outlook on potential future research. First of all, we want to underline that, considering the scope of this seminar, we only had limited time for our analysis. Since we solely focused on videos instead of other media, as for example books or exhibitions, our results should not be generalized to all forms of educational media. The same can be said regarding the language of our material. Given that we only considered German videos, the results cannot be generalized to other languages, countries our cultures. Moreover, since we only included seven educational videos, our qualitative analysis is not adequate enough for generalizing over all educational videos for children. Instead, our analysis is supposed to give a brief insight in the general presentation of AI in documentaries for children.

To conclude, it would be interesting to conduct further research without the limitations of our analysis to find more generalizable results. Furthermore, it would be interesting to investigate the effect that such presentation of AI in educational videos has on children. This would also put our current results in perspective, regarding the importance of this educational medium.

References

Der Künstliche Intelligenz-Check. 2020. Checker Tobi; Broadcasted on KIKA. https://www.kika.de/checker-tobi/sendungen/sendung126982.html.

European Commission. Joint Research Centre. 2020. *AI Watch, historical evolution of artificial intelligence: analysis of the three main paradigm shifts in AI.* eng. LU: Publications Office. https://data.europa.eu/doi/10.2760/801580 (28 March, 2021).

Gibbs, Samuel. 2017. Alphazero ai beats champion chess program after teaching itself in four hours. *The Guardian* 8.

Können Computer denken? 2020. PUR+; Broadcasted on KIKA. https://www.zdf.de/uri/ff1f26a3-8211-4cdf-abf1-c5b527293c89.

Künstliche Intelligenz. 2021. Erde an Zukunft; Broadcasted on KIKA. https://www.kika.de/erde-an-zukunft/sendungen/videos/video48066.html.

Künstliche Intelligenz – Wie schlau sind Maschinen? 2019. neuneinhalb; Broadcasted on WDR. https://kinder.wdr.de/tv/neuneinhalb/av/video-kuenstliche-intelligenz--wie-schlau-sind-maschinen-100.html.

Löwenzahn: Intelligenz auf Rädern. 2021. Löwenzahn; Broadcasted on KIKA. https://www.zdf.de/uri/55d06d88-cd58-470b-8b0e-d1f2dc31de8f.

Press, Cambridge University. 2014. *Emtional intelligence.* https://dictionary.cambridge.org/de/worterbuch/englisch/emotional-intelligence.

Studio im Netz München. 2020. *Künstliche Intelligenz – kindgerecht erklärt.* https://www.bpb.de/mediathek/301948/kuenstliche-intelligenz-kindgerecht-erklaert.

Was ist "Künstliche Intelligenz"? - logo! erklärt. 2018. logo!; Broadcasted on KIKA. https://www.youtube.com/watch?v=unAdsyOZB9c.

You, Chengcheng. 2020. The Necessity of an Anthropomorphic Approach to Children's Literature. en. *Children's Literature in Education.* DOI: 10.1007/s10583-020-09409-6. https://doi.org/10.1007/s10583-020-09409-6 (27 March, 2021).

Chapter 16

How is AI portrayed in Netflix' documentary "The Social Dilemma" and how do newspapers react to it?

Micaela Barkmann, Dana Dix & Kai Dönnebrink

In 2020, Netflix released a documentary called *The Social Dilemma*. Former employees of tech companies, like Google or Facebook, state the societal damage caused by the usage of artificial intelligence (AI) in social media. The documentary was seen by millions of viewers and discussed publicly by newspapers, for instance. Using qualitative content analysis, we investigated whether the portrayal of AI usage in social media matches the tone of the public discourse disseminated via newspapers. It was analyzed and compared both the criticism from the Netflix documentary and reviews in newspaper articles. For this purpose, the documentary's aspects were summarized in categories. Statements from the newspaper articles were divided into pro and contra arguments for the documentary itself on the one hand and AI in social media on the other. The comparison between the statements made in the documentary and the newspaper articles has shown that the perception towards social media and the AI-based algorithms used therein does not necessarily coincide with journalistic opinion. This result deviates from our initial hypothesis. Interestingly, regional differences in the tone of reporting also occurred among the examined articles.

Keywords: Social Media | AI | The Social Dilemma | Documentary | Newspaper | Netflix

1 Introduction

Netflix is a global entertainment service that provides TV series, documentaries, and feature films to approximately 204 million paid memberships in over 190 countries (Netflix 2020a, 2021). Figure 1 shows that the number of paid memberships has grown steadily in recent years in all four regions in which Netflix operates. At the end of 2020, the two largest regions were the United States and Canada, with approximately 73.9 million paid memberships, and Europe, Middle East, and Africa, with approximately 66.6 million paid memberships.

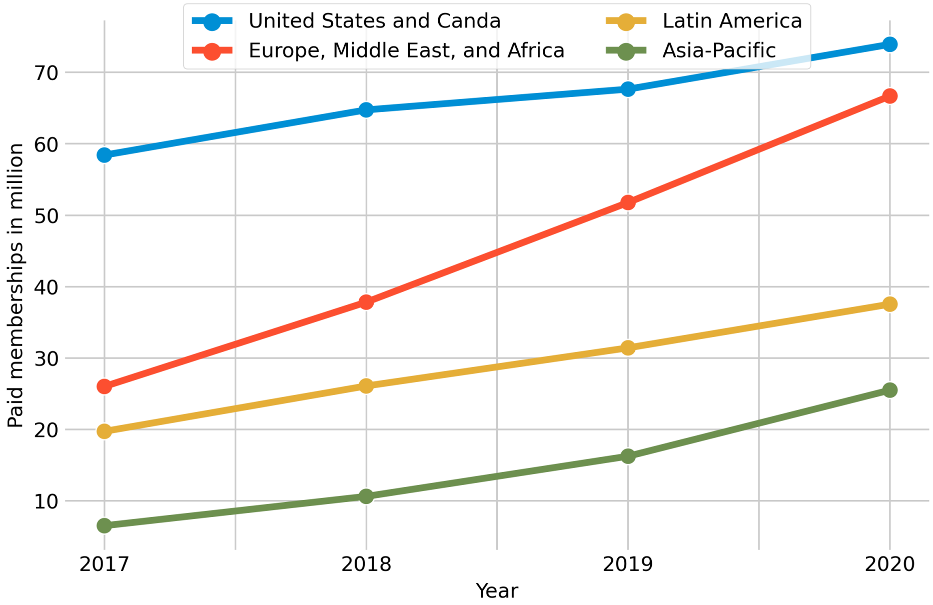


Figure 1: Paid memberships of Netflix at the end of the year in different regions (Netflix 2020a, 2021).

With this amount of paid memberships and potentially even more viewers, Netflix has a significant impact on public discourse. Furthermore, similar to traditional media, such as newspapers or TV stations, Netflix can influence the public discourse through its content. One documentary that has caused an immense public discourse is the Netflix Original *The Social Dilemma*. According to Netflix, it has been one of the most-watched documentaries in 2020 (Netflix 2020b). *The Social Dilemma* explores the rise of social media powered by AI and its consequences. Thereby, it significantly influences the public discourse about social media and the AI behind it. In response to the documentary, many major publishers have published articles on the content of the documentary. We hypothesize that the general perception of the journalistic articles reinforced the documentary and its content. To test our hypothesis, an analysis of the documentary script and some selected journalistic articles was carried out, and the results were then compared.

In the following, we introduce social media and AI's role in it, which is followed by our methodology. In the analysis, we start with analyzing the picture of AI in the context of social media drawn by the documentary. In a second step, we examine related newspapers from Germany, Great Britain, and the United States for the reactions to the documentary and their image of AI in the context of social media. Afterwards, we compare the findings of both qualitative analyses, report our results, and conclude with a discussion.

2 Social Media and AI

As *artificial intelligence*, *social media* is a widely used term in many contexts for which there seems to be no rigid definition. This may, among other things, be due to a large number of possible applications and correspondingly different characteristics, but it may also be due to the rapid ongoing development. Social media are usually described as websites or computer programs, but in any case, they are communication tools that allow users to interact online. This form of media allows users to share and consume information of various kinds (Press 2021). Depending on the platform, the content mainly consists of text, photos, and videos. Interaction occurs, for instance, through commenting, liking, and sharing posts or sending direct messages. Social media can be distinguished from traditional mass media not only by the fact that they are web-based but above all by user-generated content (Obar & Wildman 2015).

The fact that the number of social media users worldwide is estimated to rise to three billion in 2021 illustrates the impact social media can have. Facebook alone is used by 1.85 billion people every day. Since 2012, daily social media usage has increased steadily, reaching 2.5 hours per day in 2018. Germans spend an average of 84 minutes a day on social media, with 16- to 24-year-olds making up the largest group of users in Germany. 89 percent of them are actively engaged in social media (Statista Research Department 2020).

In social media, a wide range of AI algorithms and methods are used, e.g., text is often automatically translated through machine translation into the user's language, and face recognition is used to detect images of users. Besides, AI, especially the subfield of machine learning (ML), is used to personalize what is displayed to a user. In Facebook's News Feed, ML is used to determine which content should be displayed to the user. The content has to be relevant and interesting and is highly dependent on the user. According to Facebook, they "use

ML to predict which content will matter most to each person to support a more engaging and positive experience" (Akos Lada 2021). A wide range of data is used to achieve this, including general information like time, the users' personal interaction with similar content, and the interaction of different users with the specific content. Thereby, the interaction does not have to be direct, e.g. liking, commenting, or sharing a post, but can also be an indirect interaction like the user's scrolling speed, for example.

Furthermore, Facebook uses ML for personalized advertising, which works similar to the News Feed ML. However, instead of looking for the most relevant content for a user, they try to "predict a particular person's likelihood of taking the advertiser's desired action" (Facebook 2020). They take several factors like the users' interests, the ad content, and the interaction between users and ad into account. In addition, they use ML to predict an ad's quality score considering the users' direct or indirect feedback to an ad and other quality criteria for low-quality ads like sensationalized language. Both predicted values and the advertiser's bid, i.e., what the advertiser is willing to pay, are then used to calculate the ad's total value score.

3 Methodology

The base material for the analysis is, on the one hand, the transcript of the Netflix documentary (from the Loft 2020), on the other hand, a selection of newspaper articles that refer to it.

To cover the broad public discourse in the analysis, specific criteria were taken into account in selecting newspaper articles: The publishers of the selected newspaper articles had to meet a minimum average paid circulation (table 1) of one hundred thousand newspapers sold per week. This value does not include publishers' content that can be viewed online, as these values are complicated to determine. Also, it can be assumed that more than one person reads physically sold newspapers. Another selection criterion is the language used in the articles. Only articles in English and German are included in the analysis. However, no publication window has been specified since the newspapers' reactions appeared immediately after the publication of the documentary.

The selection of newspapers was limited to ten, so that material for the analysis is available in the form of ten articles from ten different newspapers. Four of the articles come from the German newspapers *DIE ZEIT* (Laaff 2020), the *Frankfurter Allgemeine Zeitung* (FAZ) (Bähr 2020), the *Süddeutsche Zeitung* (SZ) (Hurtz 2020) and *DER TAGESSPIEGEL* (Bickelmann 2020). Four more articles are

Table 1: Newspaper circulation

Publisher	Circulation in 2019	Origin / Language
The New York Times	443,000 (Watson 2021)	United States / English
Los Angeles Times	417,936 (Research 2019)	United States / English
DIE ZEIT	528,706 (4th quarter) (IVW 2021b)	Germany / German
Frankfurter Allgemeine Zeitung	241,227 (4th quarter) (IVW 2021c)	Germany / German
Süddeutsche Zeitung	337,906 (4th quarter) (IVW 2021d)	Germany / German
DER TAGESSPIEGEL	120,763 (4th quarter) (IVW 2021a)	Germany / German
The Guardian	141,160 (January) (Mayhew 2019)	Great Britain / English
Metro	1,426,050 (January) (Mayhew 2019)	Great Britain / English
The Sun	1,410,896 (January) (Mayhew 2019)	Great Britain / English
Daily Mail	1,246,568 (January) (Mayhew 2019)	Great Britain / English

from the British newspapers *The Guardian* (Naughton 2020) the *Metro* (Woodcock 2020), *The Sun* (Bellotti & Knox 2020), and the *Daily Mail* (Rhodes 2020). In addition, two articles from the American newspapers, *The New York Times* (NYT) (Girish 2020) and the *Los Angeles Times* (LA Times) (Crust 2020), were analyzed. In the following, the titles of the newspapers rather than the authors are given as sources since this is of greater importance for our context.

In this paper, a qualitative content analysis according to Mayring (Mayring 2014) was conducted. The concrete content-analytical method chosen is structuring. The aim was to extract a structure in the form of categories. Since the ontology (see chapter 2) underlying this book did not apply to the material in all respects, the categories were assigned inductively rather than deductively. All relevant text passages were extracted and grouped into corresponding categories resulting from the material. Since the documentary script is used as a starting point in this work, the category system was created based on it. Additionally, different tags were assigned for aspects within a category. The section 7 contains a list of categories, associated tags, a description of these tags, and anchor samples as prototypical examples for the respective tags.

Table 2: Tags used for the newspaper analysis

Tag	Anchor example
in favor of documentary	"All this is true. And it is not new. But it has never been conveyed as forcefully as Orlowski does."(DER TAGESSPIEGEL)
against documentary	”But the grab bag of personal and political solutions they present in the film confuses two distinct targets of critique: the technology that causes destructive behaviors and the culture of unchecked capitalism that produces it." (NYT)
in favor of AI	”And we reproduce - as much as this comparison is lame, must be lame - very similar concerns that are so often raised when something is new, no matter whether it's printing or cars, television or headphones: we fear that people will degenerate, become overstimulated and overwhelmed, we fear that they will fall into addiction and apathy, become socially impoverished. And what we actually mean is: help, we are a little overwhelmed by all the new things that have effects we cannot yet assess." (DIE ZEIT)
against AI	"It is true that social media collect millions of millions of data points about us, know much more know much more about us than we might than we might realise." (DIE ZEIT)

Since the newspaper articles can be understood as a kind of review of the documentary, more general tags, less related to specific topics, were assigned here in a first attempt to evaluate the material for positive and negative criticisms related to the documentary itself and AI. Table 2 gives an overview of used tags and corresponding anchor examples.

Sentences, or at least several consecutive words, were generally used as analysis units to capture the context adequately. In this way, a neutral, close-to-the-subject representation of the material without distortions due to pre-assumptions should be made possible. The tagging was done by searching all text material for basically relevant and meaningful text passages in the first pass. These were first highlighted. Then these text passages were collected in a new document and grouped thematically in the next step.

4 Analysis

To compare the documentary and the newspaper articles qualitatively, a content analysis of the documentary script in section 4.1 and an analysis of the newspaper articles in section 4.2 were carried out. The results of the analysis were then compared in section 4.3.

4.1 The Social Dilemma

The documentary *The Social Dilemma* investigates the current state, effects, and problems of social media. Central parts are the interviews with former employees and executives from companies like Google, Facebook, and Twitter, as well as other IT professionals. In these interviews, they stated that although social media platforms have benefits, they nowadays target the wrong goals. These goals are the *engagement goal*, i.e., how to drive up the user's usage, the *growth goal*, i.e., how to keep the user coming back and invite all of their friends, and the *advertising goal*, i.e., how the company can earn as much money as possible from advertising. To achieve all of these goals, the companies use powerful AI algorithms and psychological tricks, e.g., infinite scrolling feeds and content refreshing on a user's request that always shows new content, which is similar to how slot machines work. The main problems or symptoms that can be observed in public caused by following these goals are, for example, addiction to social media, polarization, spreading of misinformation, and massive data collections.

Alongside the interviews, the producers staged a drama. It is supposed to show the impact of social media use on young people and how the AI algorithms behind social media work. Ben, the main character of the drama, is manipulated by AI to change his behavior. In the beginning, he is a teenager that has friends and is active in his sports team but has already a high screen time for a teenager. Throughout the drama, he distances himself more and more from his friends, family, and teammates, believes in fake news, and spends most of his time in front of his smartphone. In the course of the drama, the viewer can see several times how the three AIs — played by a single person wearing different colored t-shirts — try to achieve their different goals (*engagement*, *growth*, and *advertisement*). To do this, they watch every action Ben takes, such as his scrolling speed and how long he looks at a particular image, combine them with other information, like his current location and who is nearby. They then evaluate the outcome of different actions they could take, e.g., showing a picture of a friend, personalized advertisement, or a trending video. Thereby, the AI is not interested in whether the displayed images and videos are good for Ben, but only whether he stays on the screen or not.

When talking about AI in the interviews, the experts mainly refer to AI as ML or, more specific, as deep learning due to the high amount of data the algorithms process. However, sometimes they also talk about AI in the context of Snapchat filters, for example, which then is related to augmented reality. They state that AI is not objective because it is an algorithm optimized to some definition of success, which is not objective in most cases. So, often AI is not designed to give the user really what they want to have but something that maximizes the outcome of the AI's goal. Furthermore, they claim that the brought public cannot understand the AIs, and even top experts do not understand everything that happens inside the AI algorithms. Thus, the companies have not full control over their systems. Nevertheless, they state that AI is simultaneously a utopia and dystopia.

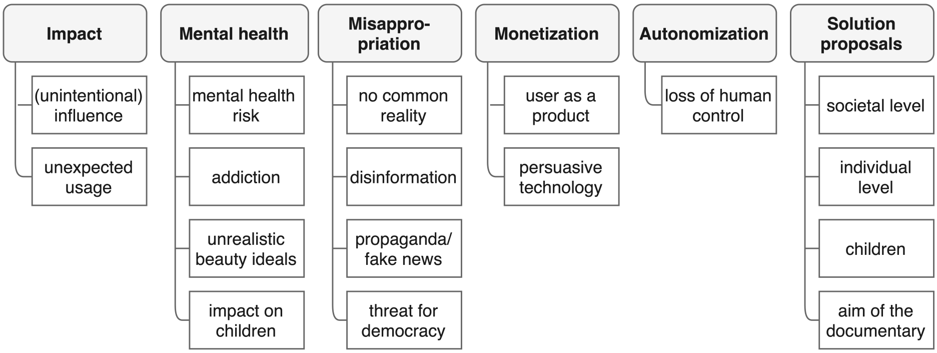


Figure 2: Inductively generated category system for the content of the Netflix documentary *The Social Dilemma.*

The category system created for the content of the documentary (see Figure 2) contains only negative aspects of social media and the use of AI in this context. Positive aspects such as “the fact that these tools actually have created some wonderful things in the world … reunited lost family members [and] … found organ donors” (from the Loft 2020) were mentioned in only two places. For this reason, they do not find any further consideration in the category system, as the content focus is clearly on the negative effects on society.

Impact, the first category, is made up of statements about how significant the influence of a few people in the tech industry is on society. Not all of this is actually intended, i.e., the actual use of developments often turns out differently than expected. The three categories, *Mental health*, *Misappropriation*, and *Monetization*, include the most detailed explanations relate to specific problems that arise from the intensive use of social media - partly as an unintended side effect, in some cases also with full intention.

Mental health, as the title suggests, deals with problems and negative effects in this area. Specific illnesses such as depression and anxiety are mentioned, but also the negative tendencies concerning self-worth and identity. Unrealistic ideas of beauty are spread through filters, for example. Addiction is another major factor that affects all age groups. However, in particular, children and younger generations suffer from the side effects of intensive social media use.

The category *Misappropriation* describes the actual intentional misuse of social media. Especially factors such as the lack of a common reality in the age of disinformation as well as deliberate manipulation, propaganda, and the spreading of fake news are a threat for democracies.

Monetization, as another category, includes descriptions of social media as a so-called "money machine", in which the user is the actual product with whose data as much money as possible is to be made. Persuasive technologies are used to change the user's behavior imperceptibly.

A series of statements are made in the category *Autonomization* about why it is so difficult to change the existing system. Here, among other things, the loss of human control plays a decisive role.

Finally, some statements can be grouped into the category of *Solution proposals* that can be understood as proposals for solving the problems mentioned before. Suggestions are made about how the tech industry and governments should participate in a change, for example, through taxes and laws. There are also concrete tips for the viewer as an individual, especially for children's social media use. In conclusion, the documentary's main goal is to inform and empower the public to engage in an open discourse to build enough pressure on governments and the tech industry.

4.2 Newspaper articles

As soon as the documentary went online, the press worldwide reacted immediately. In particular, a considerable amount of major publishing houses printed at least one article in response to the documentary. However, the reactions and opinions about the documentary were quite diverse and ranged from "most important documentary … this year" (LA Times) in a positive way to "painting a dystopia" (SZ).

While most journalists agreed that the documentary failed to present any new, surprising revelations, the *NYT*, the *LA Times*, *The Guardian*, the *Metro*, the *FAZ*, the *SZ*, and *DIE ZEIT* scored the way the information was presented as effective. According to the authors, the documentary manages to address issues such as data mining, manipulative technologies, and social media's addictiveness and conveys it to the viewer in an engaging, digestible, and in-depth way. *The Guardian* says the documentary's fictional nature is necessary because it is difficult to criticize an "industry that treats its users as lab rats". Furthermore, the documentary could explain "to the rats what is happening to them while they are continually diverted by the treats". Whereas the SZ strongly criticizes this analogy and rejects the producer's portrayal of billions of people as lab rats and degenerated zombies due to the manipulations of technology companies.

The Guardian even evokes the scenario that in a few centuries, this documentary will be the answer to what happened to "the prosperous, apparently peaceful society of the 21st century" and describes Facebook as an existential threat to democracy. Interestingly, the author of the *Daily Mail* article, Larissa Rhodes, is also one of *The Social Dilemma* producers. Thus, she writes very positively about the documentary and reinforces the statements and information contained in it. The *LA Times* and the *Daily Mail* both refer to the statement by Tristan Harris. He says in the documentary that "[i]f you're not paying for the product, then you are the product" and unreservedly supports this utterance.

Meanwhile, the*LA Times* claims that manipulating behavior through predictive AI and the collection of data was designed to create dependence and addiction.*The Sun* states that "the combination of conspiracy rabbit holes and intentionally divisive algorithms can have truly devastating consequences". The *Daily Mail* says that these companies employ teams of psychologists to help manipulate the human brain. In this context, *The Sun* cites a study in Psychological Science that allegedly showed that teenagers' dopamine levels rise when they see a high number of likes on a post.

Some of the authors also pick up on the statistics mentioned in the film regarding the rising suicide rate in the US. While *The Sun* and the *Daily Mail* endorse the statistics mentioned in the film and confirm that they are linked to the rise of social media apps, the *Daily Mail* even goes so far as to say that social media is bringing with it "a whole generation of addicted children whose self-worth and ability to connect with others may be permanently damaged". In contrast, the *NYT* and the *SZ* criticize that the documentary does not include other causes for the increase in mental illness, such as economic and social factors, but only social media use.

Further criticism comes from the *SZ*, *DIE ZEIT*, and *DER TAGESSPIEGEL*, who claim that reality claim is more complex than presented in the documentary. *The Social Dilemma* leads to hasty judgments and generalizes the effects of social media. Not every user builds their self-confidence on feedback in social media or is online non-stop (DIE ZEIT). According to *DER TAGESSPIEGEL*, even before social media, not everything was as peaceful and carefree as it is supposedly portrayed in the documentary. The *NYT* supports this utterance by writing that "with the right changes, we can salvage the good of social media without the bad". *DIE ZEIT* even quotes a paper by Bakshy et al. from 2015, according to which the most substantial effect of the "social media bubble" isn't caused by the algorithms but by the users' individual decisions. The *SZ* criticizes the documentary's one-sided nature and emphasizes that social media's technology has also enriched life. They rate it problematic that this topic is reduced to two utterances: "Social media is addictive, algorithms manipulate humanity" and that for all kinds of problems in today's society (radicalization, polarisation, addictive behavior, etc.), the internet, smartphones, and social media are to blame. The *SZ* concludes that "[t]he documentary does not show a dilemma, it paints a dystopia".

Ultimately, the *SZ* and *DER TAGESSPIEGEL* assess social media's negative criticism and the technologies behind them as "fear of the new". For example, it is compared with the introduction of the bicycle and the fact that people reacted skeptically and said that the bicycle would harm the character and make it addictive. Similar negative reactions were seen with the introduction of electricity, railroads, newspapers, radio, cars, television, etc. This rejection, however, according to the authors, is only a reaction to being overwhelmed and the fact that the concept could not be fully grasped right away.

4.3 Comparison

After analyzing the newspaper articles and the script of the documentary, a comparison was made. This was done based on the tagging mentioned above by looking point by point to see whether the respective categories were taken up in the newspapers.

Impact: The documentary discusses that only a few people influence the development of social media platforms and that there were often other intentions behind some technologies. The supposed damage that is caused nowadays was, to a large extent, not planned. In the newspaper articles, the journalists predominantly emphasize that the technologies have also enriched today's life and that not everything is bad. A "dystopia" (SZ) is described, and social media are presented as the cause of many prevailing social problems. Moreover, this doom and gloom is a symptom of skepticism towards new technologies that has already occurred several times in history (SZ, DER TAGESSPIEGEL).

Mental health: Under this category, the effects of social media on the mental health of its users, as mentioned in the documentary, were summarised. The (intensive) use of social media, according to the documentary, is responsible for increasing numbers of cases of depression and anxiety, as well as for a general individual decline in self-confidence and sense of identity in all age groups. The newspapers reacted differently to this aspect. *The Sun* and the *Daily Mail* supported these statements in their articles. In contrast, the *NYT* and the *SZ* were highly critical, as no other causes than social media (e.g., economic factors) were included in the documentary‘s description of the causes of the rising numbers of mental health problems mentioned above.

Misappropriation: Under this category, the documentary reports the (deliberate) misappropriation of social media. Originally intended to connect people, social media is now used to spread false information and propaganda, manipulate users, and dissolve a shared reality, according to the documentary. For the most part, these statements are almost uniformly supported across all articles. User behavior is supposed to be manipulated by predictive AIs (LA Times), deliberately divisive algorithms (The Sun), and teams of psychologists behind them (Daily Mail).

Monetization: This category includes content intended to point out in the documentary that social media platforms have turned the user into a product and are making the greatest possible profit with the help of the user's data. None of the newspaper articles contradict this. In the *LA Times* and the *Daily Mail*, this is even picked up approvingly and quoted verbatim from the documentary.

Autonomization: This topic is about the increasing independence of social media and its influence on people. It would be difficult to change the driving AI and the existing system. These statements were criticized in most newspaper articles. The statements were generalized (DIE ZEIT), and the reality was more complex than it was presented in the documentary (SZ, DIE ZEIT, DER TAGESSPIEGEL). Not every user is dependent on social media and affected by the consequences described (DIE ZEIT).

Solution proposals: Finally, the documentary addresses how the previously discussed problems of social media can be solved. In addition to tips on social media behavior for consumers, solutions are proposed at the institutional level (e.g., laws or taxes). Of all the newspapers, only the *NYT* implicitly addresses this aspect of the documentary by writing that social media could be positively transformed with the right changes.

5 Results and Discussion

The documentary aims to initiate a public discourse about the conditions in the field of social media use and thus put pressure on those responsible. In any case, the documentary has succeeded in encouraging public discourse, both on the social media platforms such as Twitter, Instagram, and Facebook themselves and in the journalistic media landscape. Many major publishers have taken up the topic with at least one article. However, our analysis shows that the journalists and the documentary producers do not agree on all points. This result does not confirm our hypothesis that the journalistic articles studied would share the documentary's general perception.

While the documentary is undoubtedly predominantly negative about social media, this sentiment is not reflected in all articles. While *The Sun*, *Daily Mail*, *The Guardian*, and the *LA Times* agree almost unreservedly with the documentary's content, the *SZ*, *DIE ZEIT*, the *NYT*, and *DER TAGESSPIEGEL* tend to speak out against the radical social media and AI portrayal of the documentary. Interestingly, another cautious tendency can be derived from the analysis: all German articles are critical of the documentary, while the British articles support the statements made there. On the other hand, in the American articles, one speaks in favor of the documentary and the other against it. This is a very interesting observation and would be worth further investigation to determine whether this tendency is purely coincidental or statistically relevant.

One problem that emerged during our research was that both the documentary and the newspapers talked about AI rather implicitly. Social media is addressed and its role/impact on society. However, as we have already elaborated in sections 2 and 4.1, the functions addressed in the documentary are built on AI-based algorithms and thus explicitly become part of the criticized subject matter. A possible explanation for the superficial treatment of social media's functioning could be that this is easier to grasp for the broad audience of the documentary than talking about AI and algorithms. Since the documentary has set itself the task of informing the public and encouraging public pressure for change, this approach is quite plausible and, in the context of the diverse audience addressed, probably just as effective as talking about concepts that are more difficult to grasp such as AI.

6 Conclusion

Our findings show that Netflix's documentary *The Social Dilemma* does not describe an uncontroversial social media image. The analysis of newspaper articles from some of the largest independent publishers in the United States, the United Kingdom, and Germany has shown that a different opinion is being disseminated through this public discourse channel.

Since the publication of the documentary, there have already been initial changes on the social media platforms. Whether these changes are directly related to the documentary remains to be seen. At the end of January, Facebook and Mark Zuckerberg announced that users would no longer be recommended political content. By then, a function had already been introduced to turn off political advertising completely individually, and initial tests had been run in which users were shown less political content than before (Gupta 2021). In addition, in January 2021, Donald Trump's accounts were suspended from Twitter and Facebook, respectively, "due to the risk of further incitement of violence" (Twitter 2021, Schuler 2021)). Trump's statements on these platforms are said to violate the guidelines applicable there and led to such measures.

Further research could include the addressed companies for a uniformly representative analysis. For example, a comparison could be made of how AI's use on social media is described and whether potential negative consequences are mentioned at all. It could also be examined to what extent the large tech companies react to the documentary's criticism.

In conclusion, no unified opinion on social media can be identified. However, presumably, both the documentary and the newspaper articles can be traced back to one core essence: AI-based functionalities on these platforms are designed to entice users into increased interaction. Therefore, every user of such services should be aware of the algorithms behind them and reflect on their own usage behavior. Irrespective of this, the discourse described in this article helps to make users aware of potential problems. This, in turn, also leads to the fact that responsible parties (companies and government) have to take a stand and give in to social pressure.

7 Appendix

In the following, the category system set up to map the content of the Netflix documentary *The Social Dilemma* is presented in detail. Tags were assigned to each of the six categories described in more detail and supported by a selection of anchor examples. The anchor examples are quotes taken from the script of the documentary (from the Loft 2020).

Table 3: The tags of the category impact.

Tag	Description	Anchor example
(unintentional) influence	A few people have a significant influence on society, the negative consequences of which were not necessarily intended.	”never before in history have 50 designers — 20- to 35-year-old white guys in California — made decisions that would have an impact on two billion people”
unexpected usage	Functions are sometimes used differently than expected during development.	”they take on a life of their own. And how they're used is pretty different than how you expected.”

Table 4: The tags of the category monetization

Tag	Description	Anchor example
user as product	Social media users are described as a product, as all their activities are tracked, and their attention is directed to make as much money as possible, especially with advertising.	”Many people call this surveillance capitalism, capitalism profiting off of the infinite tracking of everywhere everyone goes by large technology companies whose business model is to make sure that advertisers are as successful as possible.”
persuasive technology	The behavior of users is to be modified to the advantage of tech companies with the help of psychological methods.	”we want to psychologically figure out how to manipulate you as fast as possible and then give you back that dopamine hit”

Table 5: The tags of the category mental health.

Tag	Description	Anchor example
mental health risk	The (intensive) use of social media endangers the mental health of many people.	"A whole generation is more anxious, more fragile, more depressed."
addiction	Addiction is one of the biggest problems in this context.	"Tens of millions of Americans are hopelessly addicted to their electronic devices."
unrealistic beauty ideals	Social media conveys an unnatural image of beauty.	"These cosmetic procedures are becoming so popular with teens, plastic surgeons have coined a new syndrome for it, 'Snapchat dysmorphia', with young patients wanting surgery so they can look more like they do in filtered selfies."
impact on children	There is now a whole generation of young people for whom online connections are more important than ever before and thus have great influence.	"We're training and conditioning a whole new generation of people ... that when we are uncomfortable or lonely or uncertain or afraid, we have a digital pacifier for ourselves that is kind of atrophying our own ability to deal with that."

Table 6: The tags of the category autonomization

Tag	Description	Anchor example
loss of human control	There is hardly any human supervision, and in most cases, it is no longer comprehensible even for developers how the systems work.	"So, imagine you're on Facebook ... and you're effectively playing against this artificial intelligence that knows everything about you, can anticipate your next move, and you know literally nothing about it, except that there are cat videos and birthdays on it. That's not a fair fight."

Table 7: The tags of the category misappropriation

Tag	Description	Anchor example
no common reality	The fact that the content displayed in social media is personalized leads to different realities.	”And then you look over at the other side, and you start to think, ’How can those people be so stupid? Look at all of this information that I'm constantly seeing. How are they not seeing that same information?' And the answer is, ’They're not seeing that same information.'”
disinformation	False information spreads easily and can hardly be distinguished from the truth.	”Social media amplifies exponential gossip and exponential hearsay to the point that we don't know what's true, no matter what issue we care about.”
propaganda, fake news	False information can also be deliberately spread by misusing social media for manipulation purposes, propaganda, or spreading fake news.	”platforms make it possible to spread manipulative narratives with phenomenal ease, and without very much money.” ”There's a study, an MIT study, that fake news on Twitter spreads six times faster than true news.”
threat for democracy	Due to disinformation and the lack of a common reality, there is an increasing polarization within society, which destabilizes democracy, among other things.	”Imagine a world where no one believes anything true. Everyone believes the government's lying to them. Everything is a conspiracy theory. ’I shouldn't trust anyone. I hate the other side.' That's where all this is heading.”

Table 8: The tags of the category solution proposals

Tag	Description	Anchor example
societal level	Proposals for action are formulated here, aimed primarily at the tech industry and governments.	”We can demand that these products be designed humanely. We can demand to not be treated as an extractable resource.” ”We could tax data collection and processing”
individual level	Tips are given on how each individual can use social media more consciously or even avoid it.	”Reduce the number of notifications you get.”
		”follow people on Twitter that [you] disagree with”
children	Agreements with parents should help children in the appropriate use of their devices and social media.	”the first rule is all devices out of the bedroom ... half an hour before bedtime ... The second rule is no social media until high school ... And the third rule is work out a time budget with your kid.”
aim of the documentary	The documentary aims to inform the public and encourage discourse to build pressure, especially on the tech industry, for necessary changes.	”I feel like we’re on the fast track to dystopia, and it’s gonna take a miracle to get us out of it. And that miracle is, of course, collective will.”

References

Akos Lada, Tak Yan, Meihong Wang. 2021. *How machine learning powers Facebook's news feed ranking algorithm.* https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/ (31 March, 2021).

Bähr, Julia. 2020. *Sind wir dieser Technologie wirklich gewachsen?* https://www.faz.net/aktuell/feuilleton/medien/die-netflix-doku-the-social-dilemma-stellt-grosse-fragen-16971396.html (28 March, 2021).

Bellotti, Alex & Miranda Knox. 2020. *Netflix's The Social Dilemma reveals how chilling power of social media 'fuels suicide spikes, conspiracies & massacres'.* https://www.thesun.co.uk/tech/12754924/social-dilemma-netflix-terror-control-manipulation-facebook/ (28 March, 2021).

Bickelmann, Jonas. 2020. *„The Social Dilemma" auf Netflix: Warum diese Doku Teil des Problems ist.* https://www.tagesspiegel.de/kultur/the-social-dilemma-auf-netflix-warum-diese-doku-teil-des-problems-ist/26191666.html (28 March, 2021).

Crust, Kevin. 2020. *Review: a call to digital arms, 'The Social Dilemma' demands change.* https://www.latimes.com/entertainment-arts/movies/story/2020-09-09/review-social-dilemma-facebook-google-netflix (28 March, 2021).

Facebook, Inc. 2020. *Ihr fragt, wir antworten: Wie setzt Facebook maschinelles Lernen bei der Anzeigenauslieferung ein?* https://www.facebook.com/business/news/good-questions-real-answers-how-does-facebook-use-machine-learning-to-deliver-ads (31 March, 2021).

from the Loft, Scraps. 2020. *The Social Dilemma (2020) – transcript.* https://scrapsfromtheloft.com/2020/10/03/the-social-dilemma-movie-transcript/ (28 March, 2021).

Girish, Devika. 2020. *'The Social Dilemma' Review: unplug and run.* https://www.nytimes.com/2020/09/09/movies/the-social-dilemma-review.html (28 March, 2021).

Gupta, Aastha. 2021. *Reducing political content in news feed.* https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed/ (28 March, 2021).

Hurtz, Simon. 2020. *Social Media bedroht die Menschheit - wirklich?* https://www.sueddeutsche.de/digital/netlix-social-dilemma-kritik-1.5031070 (28 March, 2021).

IVW. 2021a. *Der Tagesspiegel (Mo-So).* with filter for quarter 4/19. https://www.ivw.de/aw/print/qa/titel/3294 (31 March, 2021).

IVW. 2021b. *Die Zeit (woe).* with filter for quarter 4/19. https://www.ivw.de/aw/print/qa/titel/967 (31 March, 2021).

IVW. 2021c. *Frankfurter Allgemeine (Mo-Sa)*. with filter for quarter 4/19. https://www.ivw.eu/aw/print/qa/titel/1056 (31 March, 2021).

IVW. 2021d. *Süddeutsche Zeitung (Mo-Sa)*. with filter for quarter 4/19. https://www.ivw.de/aw/print/qa/titel/12217 (31 March, 2021).

Laaff, Meike. 2020. *Das Dilemma mit der starken These*. https://www.zeit.de/digital/2020-10/netflix-dokumentation-das-dilemma-mit-den-soziale-medien (28 March, 2021).

Mayhew, Freddy. 2019. *National newspaper ABCs: Mail titles see slower year-on-year circulation decline as bulk sales distortion ends*. https://www.pressgazette.co.uk/national-newspaper-abcs-mail-titles-see-year-on-year-circulation-lift-as-bulk-sales-distortion-ends/ (31 March, 2021).

Mayring, Philipp. 2014. Qualitative content analysis: theoretical foundation, basic procedures and software solution.

Naughton, John. 2020. *The Social Dilemma: a wake-up call for a world drunk on dopamine?* https://www.theguardian.com/commentisfree/2020/sep/19/the-social-dilemma-a-wake-up-call-for-a-world-drunk-on-dopamine (28 March, 2021).

Netflix, Inc. 2020a. *2019 10 K-Form*. https://www.sec.gov/ix?doc=/Archives/edgar/data/0001065280/000106528020000040/form10kq419.htm (24 March, 2021).

Netflix, Inc. 2020b. *The stories that helped us escape at home*. https://about.netflix.com/en/news/what-we-watched-2020-on-netflix (24 March, 2021).

Netflix, Inc. 2021. *2020 10 K-Form*. https://www.sec.gov/ix?doc=/Archives/edgar/data/0001065280/000106528021000040/nflx-20201231.htm (24 March, 2021).

Obar, Jonathan A & Steven S Wildman. 2015. Social media definition and the governance challenge-an introduction to the special issue. *Obar, JA and Wildman, S.(2015). Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy* 39(9). 745–750.

Press, Cambridge University. 2021. *Social media | definition in the Cambridge English Dictionary*. https://dictionary.cambridge.org/us/dictionary/english/social-media (31 March, 2021).

Research, Cision Media. 2019. *Top 10 U.S. daily newspapers*. https://www.cision.com/2019/01/top-ten-us-daily-newspapers/ (31 March, 2021).

Rhodes, Larissa. 2020. *How your mobile controls your mind: More and more women are hooked on social media. But in a must-see new documentary, the experts who helped create the tech giants reveal how it's poisoning the way we all behave*. https://www.dailymail.co.uk/femail/article-8840439/How-mobile-controls-mind.html (28 March, 2021).

Schuler, Marcus. 2021. *Trumps Facebook-Seite bleibt gesperrt.* https://www.tagesschau.de/ausland/usa-capitol-trump-facebook-twitter-103.html (28 March, 2021).

Statista Research Department. 2020. *Statistiken zum Thema Soziale Netzwerke.* https://de.statista.com/themen/1842/soziale-netzwerke/ (31 March, 2021).

Twitter, Inc. 2021. *Permanent suspension of @realdonaldtrump.* https://blog.twitter.com/en_us/topics/company/2020/suspension.html (28 March, 2021).

Watson, Amy. 2021. *Average paid and verified weekday circulation of The New York Times from 2000 to 2019.* https://www.statista.com/statistics/273503/average-paid-weekday-circulation-of-the-new-york-times/ (31 March, 2021).

Woodcock, Zara. 2020. *Netflix film The Social Dilemma fans are nervous wrecks after watching shocking documentary.* https://metro.co.uk/2020/09/14/netflix-film-social-dilemma-fans-nervous-wrecks-after-watching-shocking-documentary-13270095/ (28 March, 2021).

Chapter 17

A modern god complex - Doctor Who? An analysis of (un-)specialized German news articles on AI in medical diagnostics

Isabel Grauwelman, Cosima Oprotkowitz & Katharina Trant

In this paper we aimed to provide an overview of how artificial intelligence (AI) is portrayed in the medical field and health care. We analyzed articles from specialized and unspecialized platforms from the years 2017 to 2020, written by medical experts as well as non-experts, combining a qualitative and quantitative analysis. More specifically we wanted to illustrate the current sentiments around AI-assisted diagnostics.

For the qualitative analysis we concentrated on the mentioned *Strengths*, *Weaknesses*, *Opportunities* and *Threats* (SWOT), the *Sentiments* as well as the *Demands of Action*. To corroborate our qualitative findings, we used R for the visualization of our gathered data and to also provide a quantitative analysis.

We found that there were no remarkable differences in sentiment between the different platform specializations. The majority of articles tended to communicate a positive picture of the application of AI related to medical diagnostics. Surprisingly for us, AI was described as a tool that could have the potential to make medicine more humane. To make a beneficial application possible, weaknesses of AI, like the often mentioned problems of data security and liability, were criticized and demanded to be resolved.

Keywords: Medicine | Diagnostics | German Newspapers | Artificial Intelligence

1 Introduction

Watches and apps that aim to measure your fitness, magnetic resonance imaging systems or X-ray machines that allow to look below the body's surface, pacemakers for the heart or brain to help you stay alive - technology and health care are already intertwined and undoubtedly affect everyone, whether they are patients or work in the medical field. Hence, it is no surprise that as AI receives increasingly more public attention, its possible applications in the medical sector do so, too.

Since one of us is experienced in day-to-day health care as a nurse and we are all generally interested in this topic, we wanted to analyze how the use of AI in medical diagnostics is portrayed. Accordingly, we analyzed recent German newspaper articles, some written by medical experts and some by non-experts, some published in medically focused newspapers and some in general daily newspapers. At first, we accepted every form and definition of AI in our source material, so we did not exclude any of it a priori. With the final sources, we ended up with AI only as a software and not as a "physical" entity like robots.

Our goal was to give a general overview over this public discourse with the focuses explained in the following. For a focus on a specific application of AI, see chapter 18 which covers breast cancer detection. Similarly, chapter 11 deals solely with video interviews in which physicians talk about AI in medicine.

2 Method

In our research we aimed for a combination of a qualitative and quantitative analysis to highlight the differences and similarities of how AI is portrayed in different German news articles.

For the qualitative analysis, we looked at the *Sentiments*, *Strengths*, *Weaknesses*, *Opportunities* and *Threats* (SWOT) and the *Demands of Action* mentioned in the articles. With the quantitative analysis, we additionally wanted to give a general overview of how and if there is a difference with respect to the *Year* the articles were written in and the *SWOT*. To put it into perspective, we also paid attention to *Synonyms* used for AI, e.g. "System" or "Computer", and in what kind of articles they were used.

2.1 Sources

In order to get a variety of news articles, we decided on some keywords to search with via Google. We aimed for a broad overview, which is why we also included keywords which represent potentially negative or positive associations with AI that someone from the broad public might have:

- KI in der Medizin *(AI in medicine)*
- KI und Ärzte *(AI and physicians)*
- KI Diagnostik *(AI diagnostics)*
- KI im Gesundheitssystem *(AI in the health care system)*
- KI ersetzt Ärzte *(AI replaces physicians)*
- KI hilft Ärzten *(AI helps physicians)*

We limited our search to articles from 2017 to 2020 and only included those appearing on the first two result pages. Furthermore, we defined which platforms we considered to be "specialized" and which "unspecialized": Specialized platforms were those that exclusively report on medical topics, as well as directing their articles to an audience which is familiar with medicine and healthcare, e.g. "Ärzteblatt". Unspecialized platforms were those which neither exclusively report on medical topics, nor have another thematic focus and therefore have a wider audience, e.g. "FOCUS".

Lastly, we defined the terms "experts" and "non-experts": Experts, according to our definition, were people who work in the medical field, e.g. physicians. Thus, non-experts were journalists with no medical background.

We ended up with a total of 24 articles and in Figure 1 you can see the distribution of articles on specialized and unspecialized platforms. As one can tell there is an uneven distribution, which leads to an under-representation of the specialized platforms.

Regarding the proportional distribution of experts and non-experts among the specialized and unspecialized platforms, we encountered a problem with unknown authors in the specialized articles, as seen in Figure 2 below. We were not able to figure these out as they were not mentioned in the articles, but still kept them for our qualitative analysis, as our focus was on the platform specialization. Concurrently, we decided to exclude them in the author-related parts of our quantitative analysis because we did not want to assume that they make up a homogeneous group.

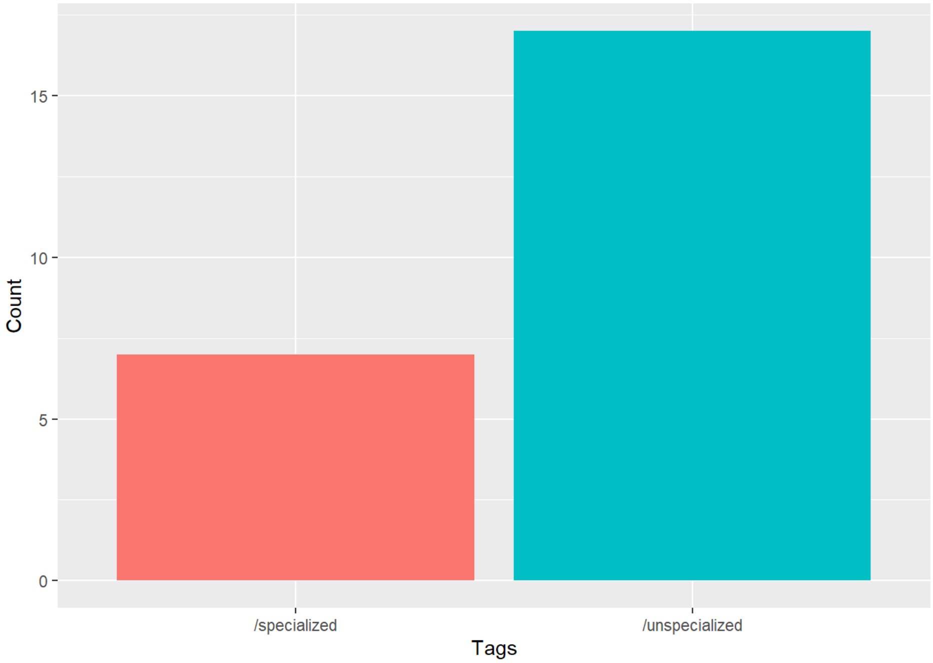


Figure 1: Distribution of article specialization

2.2 Approach

We collected our articles and uploaded them to CATMA (https://app.catma.de/catma/), a collaborative computer assisted application for text markups and analysis. To work with the articles and extract data for our analysis, we defined a tagset along which we tagged the articles. This tagset was based on the common ontology (2) and adapted to our medical context. In Table 1 you can see our final tagset with its respective definitions.

We split the articles in three parts, therefore everyone tagged a smaller portion of our collection. After that, we cross-referenced them to correct our tags and finalized the tagging process. In the end, everyone worked with every article and approved of all set tags. After this, we extracted the gathered data from CATMA and used R to visualize it for the quantitative analysis. The complete code can be accessed at https://github.com/igrauwelman/AIPD.

For the qualitative analysis we used MURAL (https://www.mural.co/) for visual brainstorming, to sort our findings into broader categories. For every part of *SWOT* and *Sentiment* as well as for the *Demands of Action* we created a subfield

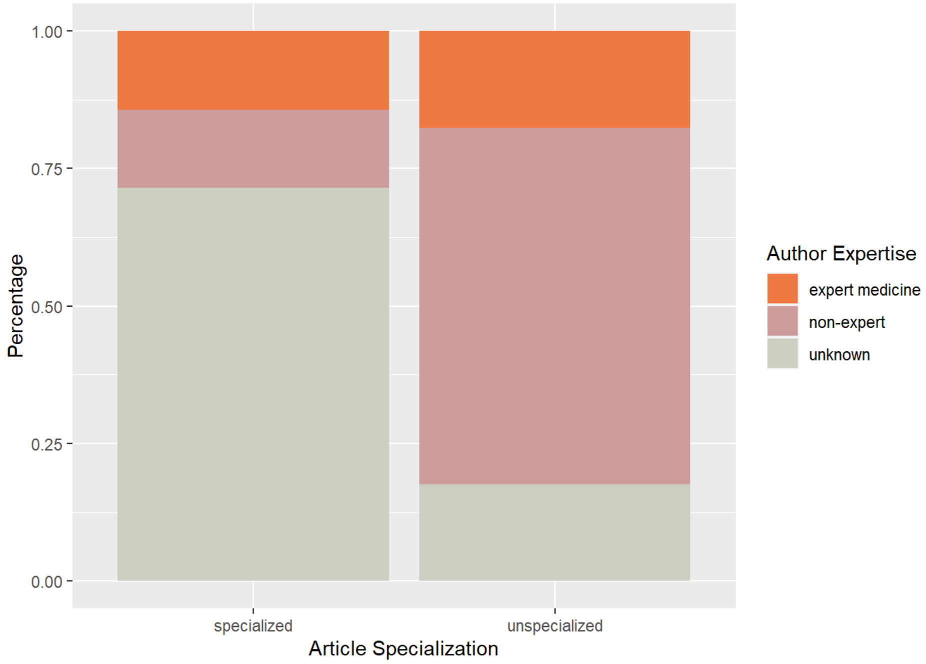


Figure 2: Specialization of the articles and authors

to collect our findings. This for example resulted in a category “Classification and Pattern Recognition” for the subfield “Strengths”.

3 Expectations

Due to our subjective belief of the sentiment of non-experts towards AI and the fact that they are not professionally involved with AI or medical diagnostics, we expected them to be generally more negative. Coherently, we thought that medical experts are more positive towards current and future applications of AI.

One thing we were sure to read repeatedly, especially among the articles written by non-experts, was the fear of AI stealing jobs and the loss of humanity in medicine. Matching this, we assumed that the dangers surrounding AI would be emphasized frequently.

Table 1: Tagset

Tag	Tagging Method	Definition
Year of publication	once	2017, 2018, 2019, 2020
Specialization	in headline and reference of author	specialized platform (expert (author)/non-expert (author)), unspecialized platform (expert (author)/non-expert (author))
Sentiment	per paragraph	contra AI, neutral, pro AI
SWOT	when mentioned	Strength, Weakness, Opportunity, Threat
Demands of Action	when mentioned	demands
Synonyms	every time when mentioned	Algorithm, Application, Big Data, Brain, Chatbot, Cloud, Computer, Deep Learning/Networks, Digitalization, Machine, Machine Learning, Product, Program, Software, Solution, System, Technology

4 Findings

In the following, we present the results of our analyses. We start with the qualitative analyses of *Sentiment*, *SWOT* and *Demands of Action*, followed by the quantitative analysis of the *Synonyms*. The section for the *SWOTs* starts with a quantitative analysis as well.

4.1 Sentiment

This section focuses on the aspects that were positively or respectively negatively reported on in the articles. We analyzed the sentiment with regard to the use of language and general tone. For example, *"With the touch of a button, AI could load all relevant publications onto the physician's screen or point him towards new research findings [...]"* (translation of Witte 2019a) is a positive example, while *"It makes it clear that physicians are not to be replaced for the time being - because* *a dermatologist has to be able to classify multiple skin alterations, [...] not exclusively differentiate black skin cancer and birthmarks."* (translation of Buck 2019) is a negative example.

There is an inevitable overlap with the SWOT section further down, as strengths and opportunities tended to be reported more positively, while weaknesses and threats tended to be reported more negatively. Therefore, the following sections just give a broad overview of the sentiment to avoid too many double-mentions. There was no notable difference regarding the platform specialization, which is why they are not differentiated in the following. As seen in Table 1, we also had a sentiment tag "neutral", which we discovered later to rather be a placeholder for sections that for example just list facts. Accordingly, we chose to not analyze this sentiment since it is quite uninformative.

4.2 Pro AI

The articles reported positively about AI as a medical application with regard to AI easing the work for physicians and thereby the treatments for patients.

4.2.1 AI as a tool for the physicians

Many articles positively emphasized AI as an assistant which does 'tiring' routine work like checking medical images for illnesses – a task that takes a lot of time and concentration (e.g. Stratmann 2020, Redaktionsnetzwerk Deutschland 2020). As a result, the physicians could concentrate more on the difficult cases and at best make fewer mistakes (e.g. Herbe 2018). For this, physicians would have to cooperate with AI, which was positively described in some articles (e.g. Healthcare in Europe 2020). On a similar note, there was positive reporting on a study that suggests that AI could also help more inexperienced physicians to diagnose on the same level as experienced physicians (Healthcare in Europe 2020).

Noticeably, when the articles reported positively, AI was mostly treated as a tool for physicians rather than their replacement (e.g. Matera 2020). As such, AI was said to for example be able to help in (literature) research, which would be especially helpful because of the increasing amount of available medical data presumably making it harder for human physicians to process all new information (Till 2020).

Another reported 'tool-like' application would be to let the AI function as a kind of 'second eye' to countercheck a diagnosis (Röhrlich 2020). This 'second eye' could also be used to gain insight into kinds of patterns that otherwise would probably be undetected, e.g. the analysis of cancer registries to detect region-based clusters (Redaktionsnetzwerk Deutschland 2020).

4.2.2 Improving the health system

Besides AI reportedly bringing more precise diagnoses (strz/Esanum 2018), it was reported that it - contrary to prior expectations - could turn medicine more humane: If physicians had to spend less time with diagnostics and administrative work, they could ideally spend more time with their patients and accordingly be able to focus more on empathy and communication (e.g. Rabhansl & Maté 2020).

Overall, the articles tended to be positive about the potential of faster and better diagnostics (e.g. Till 2020), also in regard to the costs that could be saved by AI, as it might make for fewer surgeries and examinations (Kuhn et al. 2018). Generally, this was reported to make the health system more efficient (e.g. Buck 2019), which was also reported positively.

4.2.3 Advanced qualities

Another application positively brought up was the location-independent availability of knowledge - provided that AI is internet-connected and internet is available at the respective location (Redaktionsnetzwerk Deutschland 2020) - which would be particularly convenient in regions where medical structures are weak and specialists are rare (ZEIT online 2019).

Positive reporting was also used for opportunities that are very distinct from today's possibilities: Recognizing yet unknown patterns like mentioned above or the possibility to produce digital twins of organs, which could help physicians derive patient-specific properties like the pumping capacity of the heart, could potentially change how diagnostics is done (Buck 2019).

Lastly, there were positive reports about cases where AI outshone humans, i.e. where AI was better at diagnostics, following the prospect of fewer mistakes in future diagnostics (Herbe 2018). A very prominent example was the case of black skin cancer from Heidelberg, where dermatologists and an AI competed against each other in analysing and correctly diagnosing 100 images of skin alterations. The AI recognized 95% of the cancer images correctly, while the physicians only recognized less than 90% correctly (Kröplin 2018).

4.3 Contra AI

The aspects that were negatively reported on mostly concerned the limitations and danger of AI, as well as missing regulations regarding its application.

4.3.1 Technical limitations

Firstly, it was negatively discussed that medical knowledge and experience could not be modelled accurately by an AI (e.g. Buck 2019) and that AI lacked transparency in regard to the results (e.g. Matera 2020).

Moreover, the articles tended to negatively describe cases where AI could not do what it was not trained for. For example, the AI from Heidelberg that was trained to detect black skin cancer was criticized because of its inability to detect and differentiate white skin cancer and benign skin birthmarks (e.g. Redaktionsnetzwerk Deutschland 2020).

Further, faulty results, i.e. false positives and false negatives, were unsurprisingly negatively emphasized, often linked to the note that AI assistants could do harm, too (Witte 2019b).

4.3.2 Dangers of AI application

When dangers concerning AI were mentioned, they were often related to humans' reactions: On the one side, some patients might be easily irritated or scared by a physician-independent diagnosis, for example through an AI-based app (Kuhn et al. 2018). On the other side, there could be physicians who put too much trust in AI and thus compromise their diagnoses (Redaktionsnetzwerk Deutschland 2020). In one article, the concern was brought up that many risky, unnecessary and costly examinations might be needed when an AI for instance proposes many different possible illnesses and physicians would then want to test for all of them (Kuhn et al. 2018).

4.3.3 Missing regulations

Concerning this point, in many articles the lack of standardized regulations concerning studies about AI was critically addressed (e.g. Redaktionsnetzwerk Deutschland 2020).

One article also sceptically mentioned that big corporations like Google may take over the market of medical AI due to their ability to gather personalized data and therewith develop better algorithms, which could easily clash with (national) personal data privacy regulations. The lack of regulations concerning this potential data misuse was therefore criticized (Röhrlich 2020, Kuhn et al. 2018).

4.4 SWOT

In this section, we go more into detail regarding the addressed strengths, opportunities, weaknesses and threats. As seen in Figure 3, the only difference between

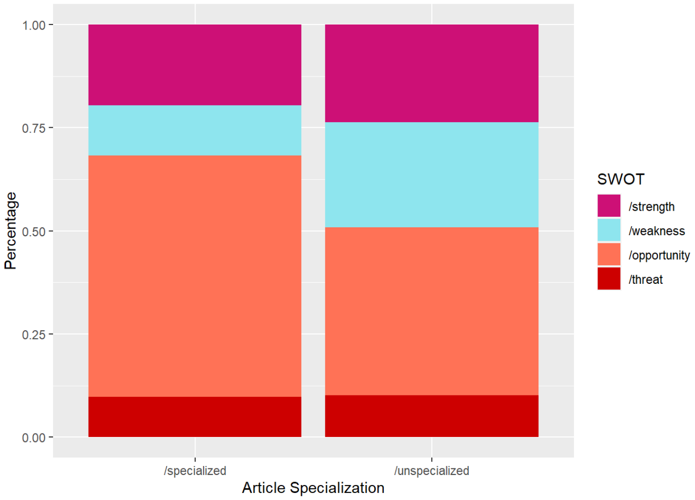


Figure 3: Article Specialization and SWOT

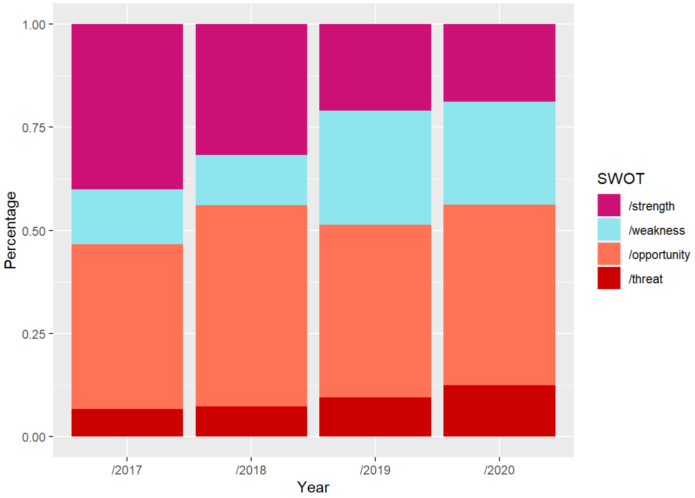


Figure 4: Year and SWOT

specialized and unspecialized articles regarding the discussion of SWOTs lied within specialized articles mentioning more opportunities and less weaknesses than unspecialized articles, whereas strengths and threats were equally often addressed. Over the years, a linear development could be made out for the mentions of strengths and threats (see Figure 4): With every year, noticeably less strengths were pointed out, while there was a small increase in the discussion of the threats. The opportunities in turn remained a rather consistent topic, whereas the number of the indications of weaknesses increased distinctly in articles written in 2019 and 2020 compared to those written earlier.

4.5 Strengths

The reported strengths mostly concerned technical characteristics of AI and which benefits it would bring to physicians in diagnostics. There was no remarkable difference between specialized and unspecialized articles.

4.5.1 AI characteristics

Some reported strengths stemmed from AI's properties, i.e. it could be faster than humans and powerful when working on a lot of data (e.g. Beeger 2017). AI's ability to learn continuously was another strength pointed out regularly, as new information was continually available (Matera 2020).

In some articles, mostly from 2019 and 2020, AI's strength concerning classification and pattern recognition was also highlighted (e.g. Rabhansl & Maté 2020) as potentially helpful in image processing (Stratmann 2020).

4.5.2 AI & the physicians

Concerning AI in relation to physicians and patients, prominent reports were about cases or studies where AI in some way "defeated" humans or scored better at diagnoses. An oftentimes, also in this chapter, mentioned example was the study on detecting black skin cancer in images which was mentioned at least once per year (Redaktionsnetzwerk Deutschland 2020, Till 2020, Buck 2019, Kröplin 2018).

Less extreme, but similarly, some articles reported how AI could "correct" human limits or mistakes, e.g. for visual analyses (e.g. Röhrlich 2020). As already pointed out in section 4.2.1, this could possibly lead to less false and more precise diagnoses (e.g. Till 2020) or to the recognition of patterns the human eye is blind for, e.g. to classify and distinguish the different characteristics of schizophrenia (Forschungszentrum Jülich 2019). Some articles also emphasized the benefits of faster and improved diagnostics and better treatment from the perspective of the patients (e.g. rme/aerzteblatt.de 2017).

4.6 Opportunities

This section deals with the opportunities AI could reportedly bring to diagnostics. As those opportunities mainly arise from strengths, some aspects are similar to the section above.

4.6.1 AI characteristics

When specific characteristics of AI were spoken of as opportunities in the unspecialized articles, they were mostly mentioned as “tools” (Till 2020). These could deal with complex or big data sets for automatic pattern recognition and classification, for example to recognize whether an abnormality on an image is benign or indeed cancer, which could take much repetitive work off of physicians (Witte 2019b).

4.6.2 AI as medical support

Similar to the strengths, the relation between humans and AI was reported to be full of potential. Very consistently mentioned was the opportunity of AI assisting physicians to help with diagnoses (e.g. Kröplin 2018), for example to preselect cases for physicians as previously mentioned (Focus Online 2019) or judge a case’s urgency (Kuhn et al. 2018). In this context, AI was reported to “adjust” or step in where human limits were reached, be it to find very small tumors (rme/aerzteblatt.de 2017) or to detect pain that patients would not always be able to express properly (Herbe 2018).

Further, an opportunity following early and better detection of illnesses was to save patients (and physicians) from unnecessary, often uncomfortable and costly examinations and surgeries (e.g. Kuhn et al. 2018) and simultaneously improve therapies such that the patients’ health would be affected positively (Redaktionsnetzwerk Deutschland 2020). On the same note, AI could detect who would benefit from a medically or financially expensive procedure beforehand, also ensuring a better treatment (Witte 2019a). Here, we were able to see a small difference through the years, as articles from 2017 and 2018 rather focused on harder to diagnose, i.e. rare or psychological, illnesses, while articles from 2019 and 2020 stated this opportunity more generally, regardless of the illness (e.g. Herbe 2018, Buck 2019).

Additionally reported were new approaches in the education and training of prospective physicians (e.g. Rabhansl & Maté 2020), which occurred more often in 2019 and 2020. The articles here saw an opportunity for medicine to become more humane (e.g. Kröplin 2018: see also section 4.2.2).

On a related note, AI was reported to help democratize medical knowledge, by making it more openly accessible, for example through apps (Röhrlich 2020). It could then also be accessed location-independently (Kuhn et al. 2018, ZEIT online 2019: see also section 4.2.3).

4.7 Weaknesses

Besides the positive aspects, the articles also addressed critical elements of AI that have to be considered. Overall, unspecialized articles contained much more aspects of AI's weaknesses than specialized articles. The focus in specialized articles was more on the shortage of research in the area and the remaining uncertainties surrounding the implementation of AI, while the unspecialized articles instead rather covered the weaknesses of AI related to its "technical" properties.

4.7.1 Technical limitations

One of these limitations of AI was that it could only do what it was trained for (Buck 2019), e.g. when trained to classify black skin cancer, it was unable to recognize white skin cancer (Kröplin 2018) and as it could not identify them, rare diseases might remain undetected (Matera 2020). Likewise, it could always just be as good as the underlying data it was trained on (Kröplin 2018, Buck 2019). The "question-answering computer system" Watson (Wikipedia Authors 2021b) for example was rather immature for its application in the health care system due to its recommendations being too trivial and of no real use for the physicians (Witte 2019a). Also, it was trained on data from American hospitals which did not make it applicable internationally, as different countries have different standards and regulations in their health systems (Witte 2019a).

For collecting larger data sets and making a wider use possible, the data gathered by different hospitals could be merged, but this turned out to be problematic because of non-uniform data that the AI could not process (Beeger 2017).

This data needed to be prepared by humans, i.e. for the AI to be able to learn certain labels, the data had to be labelled accordingly (Klöckner 2019, Redaktionsnetzwerk Deutschland 2020) which was why large data sets necessary for training were still missing (Matera 2020). Only these large data sets allowed for detecting patterns that were not visible in small data sets (Röhrlich 2020).

Another issue lied within the black box property that would make it impossible to reconstruct the decisions of the AI (e.g. Kaulen 2020). Moreover, as no system is flawless, they always would have to be controlled by humans that could intervene in case of malfunctioning (Till 2020). For example, when X-rays already carried markers made by the physicians, an AI was misguided and learned incorrect patterns (Redaktionsnetzwerk Deutschland 2020). Another flaw would be that it could also classify conspicuous cases as unremarkable and vice versa (false negatives and false positives) (e.g. Witte 2019b, Focus Online 2019). Some articles therefore concluded that AI-based analyses would actually not be (much) better than those made by human physicians, also because humans had the ability to "think outside the box" (ZEIT online 2019, Matera 2020) and the knowledge and experiences of human physicians were reportedly only possible to be modelled up to a limited extend (Witte 2019b). AI could only recommend therapies and treatments solely based on biological facts, with a supposed inability to model personal wishes and offer the emotional support patients might need (Buck 2019, Witte 2019b,c, Till 2020).

In general, these limitations were mostly mentioned in unspecialized articles with considerably more concrete aspects contained in articles from 2019 or later.

4.7.2 Dangers

Possible dangers that were not perceived yet as a threat, but could still have alarming consequences, were brought up most often alongside the limitations. The already mentioned weakness of false negatives and false positives, that was contained in at least one article every year except for 2017 (Kröplin 2018, Focus Online 2019, Redaktionsnetzwerk Deutschland 2020), could possibly lead to non-detection of dangerous diseases or unnecessary and risky medical tests (Matera 2020, Kuhn et al. 2018). Due to the nontransparent characteristic of AI, it would not be possible to compensate for errors by manual correction (Till 2020).

To make the analysis and diagnosis possible in itself, a lot of different information needed to be connected. This would increase the risk of not being able to maintain identity protection as part of the basic principles of data security (Röhrlich 2020). With the digitalization, hospitals also made themselves vulnerable to cyber attacks (Beeger 2017).

Overall, assistance systems therefore could cause a lot of harm - to give a concrete example: AI-based health apps were reportedly sometimes also used by physicians themselves to get information on possible differential diagnoses, but it was not clear whether they were aware of the limitations of said health apps or if they would rely heavily on them (Kuhn et al. 2018). One study already showed that less experienced physicians tended to generally trust AI more than experienced physicians (Healthcare in Europe 2020).

4.7.3 Further development required

Another weakness frequently discussed was the so far immature development of current AI-based systems in the health care system, as they still had to be validated and improved. To ensure the admission of an AI, the additional establishment of regulations concerning the liability issue and data security reportedly was due, considering that patients would have to provide their (anonymized) private data in order to make more individualised medicine possible in the first place (Witte 2019b,c, Beeger 2017). Non-experts did not address any aspects associated with this.

4.8 Threats

In comparison with the strengths, opportunities and weaknesses pointed out in all articles, possible or perceived threats were rather infrequently mentioned. Nevertheless, they need to be considered as well.

Expert authors that addressed threats tended to focus more on those concerning the diagnostic process, while the other authors additionally emphasized legal issues, possible changes in society and fear of replacement.

4.8.1 Legal issues

The issues concerning the legislation were only brought up in unspecialized articles. Apart from the still open question of liability - if it would even be possible to find someone to hold accountable (Witte 2019b) - the distribution of data-power seemed to have gained in importance as only articles written in 2020 dealt with this concern (Rabhansl & Maté 2020, Röhrlich 2020).

Due to the sensitivity of the patients' medical data, it would be crucial to ensure that it did not "get into the wrong hands" (Röhrlich 2020). But this reportedly might be difficult to control: Independent corporations would (need to) work with the data and such have it at their disposal to use it for other purposes that may not have been agreed on or violate data privacy regulations (Rabhansl & Maté 2020). On a similar note, a small subset of these corporations might gather significantly more data and therefore could gain more power than others. Thus, they could dominate the market which could possibly lead to a monopolization of medical data (Röhrlich 2020).

4.8.2 Risks of the application

As mentioned above (section 4.3.2), the blind trust of the physicians could be dangerous. For example, it could lead to wrong medical advice if they based it on potentially false results of an AI (Klöckner 2019). A study showed that even experienced physicians can be misled by AI and follow its advice against their own correct assessment (Healthcare in Europe 2020). Similarly, physicians that use AI-based health apps for getting differential diagnoses could arrange more and potentially risky medical tests than they would have done without the AI (Kuhn et al. 2018).

The rise of such health apps could also unsettle the patients due to information overload or misinformation, given that they would use them privately without moderation or critical questioning (Kuhn et al. 2018).

Moreover, the exaggerated reporting of study findings on AI being far better than human physicians could have major consequences for the patients' health if it hastily motivated the implementation of AI without further validation (Kaulen 2020).

It was noticeable that the articles written in 2018 purely refer to the threats coming from the rise of the AI-based health apps, whereas the risks of blind trust in AI recommendations was the primary focus in the articles written in 2019 and 2020. The articles written in 2017 did not address these threats at all.

4.8.3 Implications on future norms

In the long run, the growing usage of AI could elicit more profound changes. For instance, it could be deemed as "malpractice" if a physician did not apply AI in their diagnostic process (Buck 2019). Conversely, if patients would not want to digitize their data they could be considered "second-class" as AI could not diagnose their illnesses (Klöckner 2019). This already pointed to the eventuality of discrimination at work and in health or life insurance, disadvantaging those refusing to adjust to the digitalization (Röhrlich 2020). Further, over time, something like a three-tier health system could emerge, dividing patients into those paying for AI-based services, those not able to pay for these services and instead paying with their data and those still able to afford human physicians as they became rarer and thus more expensive (Rabhansl & Maté 2020). These threats were only mentioned in unspecialized articles written in 2019 or 2020, experts did not address any of them.

4.8.4 Fear of replacement

Since AI gained attention and started to outperform humans, the fear of replacement accompanied every new development. The medical field would be no exception and so the fear of AI replacing dermatologists, radiologists and human intelligence in general persisted. However, this aspect was only mentioned by two articles, while one of them directly included its refutation (Healthcare in Europe 2020, Kröplin 2018).

4.9 Demands of action

Following the controversy of the topic, 17 out of the 24 articles contained at least one direct demand on the future of AI in the healthcare system.

4.9.1 Algorithms, studies and research

A persistent weakness of AI reportedly was its non-transparency, also known as its "black box" property, that would lead to incomprehensible results. Because of this, in an interview, a physician pointed out that there needed to be transparent algorithms that would allow physicians to understand the reasons behind the conclusions the AI made (Witte 2019b), even though this implementation might be more complex (Till 2020).

One article in particular (Kaulen 2020) also focused on the problem of the existing studies claiming that AI was better than human physicians: They were not conducted with uniform study standards (as also mentioned in ZEIT online 2019) and not transparent enough (as also mentioned in Kröplin 2018). The researchers cited in this article called for high quality studies that would be replicable and would compare the AI with a bigger group of participating physicians which was more representative of human medical expertise. Likewise, further research was requested to overcome the uncertainties about the potential and possible threats that would come with the application of AI (Healthcare in Europe 2020).

For AI then to be applied in the diagnostic routine, the studies would have to be substantial enough to confirm the validity of AI's benefits and abilities, just like any other medical device had to be validated to be functional and beneficial (Kuhn et al. 2018, Witte 2019a). In one article, a digital expert with medical background even proposed something like a specialist examination (German: "Facharztprüfung") for AI like those human physicians had to take as well before being licensed (Buck 2019). The AI in general ought to adapt to the workflow of the physicians, not vice versa, as it was supposed to be an assistant, not a replacement (Stratmann 2020).

On a more abstract level, two expert authors noticed that the researchers that design AI systems for healthcare services did not focus enough on what the patients actually want. They demanded that the researchers considered these wishes more in designing the algorithms of AI systems and also included ethicists, patients and specialists into that process (Witte 2019a,b).

These demands were nearly only urged for in unspecialized articles, with only one specialized article addressing the need of evaluating the so far insufficiently validated AI systems. From these unspecialized articles, only those written by experts bring up the necessities regarding the further development of AI. In turn, only those written by non-experts point out the need of more and specifically more transparent studies. The articles from 2017 and 2018 covered these demands in a very abstract manner ("even more - and especially more transparent - studies" (translation of Kröplin 2018), "evaluation of these so far insufficiently validated systems" (translation of Kuhn et al. 2018), while the articles written in 2019 and 2020 contained more concrete demands, for example "high scientific study standards" (translation of Kaulen 2020), "orientation on what is important for the patients" (translation of Witte 2019a), "a kind of a specialist examination for AI systems" (translation of Buck 2019) and "algorithms as transparent as possible" (translation of Witte 2019b).

4.9.2 Safety concerns

A frequently mentioned demand was on data security and patient safety: The patient would have to share highly sensitive data in order for AI to be used in the diagnostic process. It reportedly was important that the patient could keep the sovereignty over their data and hold the right of withdrawal (Kuhn et al. 2018, Klöckner 2019, Bensch 2017).

Apart from that, the application of AI would have to be on the basis of carefully thought-through security concepts that prioritized the health and safety of the patient (Bensch 2017). For this, one article also emphasized that humans would have to remain in control over AI and frequently check for mistakes in the algorithms, because "machine learning must never be completed" (Klöckner 2019).

Overall, these concerns were only brought up by non-expert authors, both in unspecialized and specialized articles (e.g. Rabhansl & Maté 2020, Bensch 2017). While one expert author only vaguely touched upon the importance of data security (Kuhn et al. 2018), data security in general was mentioned every year in at least one article (Beeger 2017, Kuhn et al. 2018, Klöckner 2019, Röhrlich 2020), marking it as an ongoing concern.

4.9.3 Legislation and state

As stated above, one of the most often mentioned demands was the necessity of establishing legal frameworks around the medical use of AI. More specifically, the liability issue would still need to be settled to know who would be legally accountable in case of mistakes (Witte 2019b,c). Furthermore, the usage of AI might soon be possible at any time and for every patient. This raised the question if the patients would still have the choice over whether AI was used for their diagnostic process or if the physicians or the hospitals would make that decision on their own (Witte 2019b). The legislation created around the use of AI therefore had to comprise not solely prohibitions, but more importantly should give guidelines and rules about when to employ AI, where the use of sensitive patient data would be appropriate and to what extend it would be allowed to base important medical decisions on AI recommendations (Röhrlich 2020).

Another problem supposedly would be the lack of uniform digital data: each hospital still had its own system in digitizing patient data which made it a lot harder to merge it nationally and internationally to build big data bases as the foundation of AI training (Beeger 2017). National progress was demanded to organize uniform digitalization in order to overcome this difficulty (Till 2020). The implementation of the electronic health record (German: "elektronische Patientenakte"), which actually started in January 2021 (Wikipedia Authors 2021a), was said to go hand in hand with this (Witte 2019a). This data should be available for researchers to foster new and further research based on real patient data (Buck 2019). One article claimed that Germany needed to "catch up" with the digitalization of health in general, implying that Germany was too cautious to make progress possible (Klöckner 2019).

The only difference we could detect regarding these demands was that only unspecialized articles made the point of Germany's leeway in terms of digitalization in the health care system and that national progress would be due (Till 2020, Witte 2019a, Klöckner 2019, Buck 2019). In general, all of the here cited articles were written in 2019 or 2020, while one article from 2018 mentioned the need of "legal frameworks" only vaguely (Kuhn et al. 2018).

4.9.4 Requirements for the physicians

The group that supposedly is the most affected by the rise of AI implementations in the health care system are the physicians, so their collaboration would be crucial to make AI application successful (Beeger 2017).

With the increasing number of AI-based health apps alone, the physicians reportedly would need to familiarize themselves with the basic principles to understand their patients' concerns and assist them with the use of said apps (Kuhn et al. 2018).

This already pinpointed the direction of the transformation of the medical profession, which seemed to be more of an adaptation to the patients' needs: It was said that the physicians were not only required to be educated about the technology and its application (Kuhn et al. 2018), but also concentrate more on the humane dimension of their profession by focusing on empathy and conversation training to optimally support their patients (Rabhansl & Maté 2020). They reportedly would need to critically reflect their role as a physician to accept the changes in their profession that would presumably happen (Röhrlich 2020).

Besides the AI- and patient-centered demands, it was also mentioned that the physicians should get further medical training to be as good in diagnostics as AI (Redaktionsnetzwerk Deutschland 2020).

Both specialized and unspecialized articles, regardless of their authors' expertise, discussed what would be required from the physicians to adjust to and learn from the rise of AI in the health care system without any striking differences. However, there seemed to be a slight shift from 2017 and 2018 to 2020 in what was expected from the physicians: At first, they ought to understand the basic principles of health apps and be generally more open towards change induced by AI, but more recently they were asked to concentrate on their medical training and reflect on their role as a physician, i.e. be more active in adapting to the "AI revolution" (Kuhn et al. 2018, Beeger 2017, Redaktionsnetzwerk Deutschland 2020, Röhrlich 2020).

4.9.5 Public discourse and inclusion of the general population

The press shapes the public discourse and perception to a great extend, which is why it is crucial how the press presents AI and the advantages and threats it may come with. In two articles therefore there were demands that the press should report differently about AI and its usage in the health care system: The focus should be on the collaboration of human and machine and its advantages and not on the competition between them, fostering the fear of AI completely replacing physicians (Healthcare in Europe 2020). At the same time, the benefits of the implementation of AI would need to be brought to the public's attention to clear up misunderstandings, ease unjustified fears and maybe even incite enthusiasm about it (Klöckner 2019, Beeger 2017).

More directly, it would also be important to educate the public about digitalization and the basic principles of AI, without the informing press as the only linkage between science and general public, to counteract the otherwise predetermined “digital divide” (Kuhn et al. 2018). This “digital divide” describes the problem of only some members of society being knowledgeable about digitalization while others are not, leading the former to have an advantage over the latter (Cambridge Dictionary 2021). With this knowledge, they could actively take part in designing the AI systems instead of facing it passively (Witte 2019b).

Only non-experts wrote about the importance of the press and that the reporting should be more in favor of the collaboration because it influences the acceptance of the population, which one article also addressed as something that needed to be discussed (Kuhn et al. 2018). This demand was part of articles written in 2019 and 2020 (Klöckner 2019, Healthcare in Europe 2020), whereas the importance of educating the population about the technical advances and digitalization was pointed out in articles written in 2017 and 2018 (Beeger 2017, Kuhn et al. 2018). In 2019, that aspect remained only vague by prompting that the general population should “actively participate” in the development, without stating further details (Witte 2019b).

4.10 Synonyms

In total, we identified 17 distinct synonyms as a replacement for the term “AI”, some more frequent than others: As Figure 5 shows, the terms “Algorithm”, “Computer” and “System” were significantly most often used, while “Software”, “Technology”, “Machine”, “Deep Learning / Neural Networks” and “Chatbot” stood out as well.

When looking at possible tendencies of specialized and unspecialized platforms (see figure 7), some differences could be pointed out: most strikingly, the term “Computer” was significantly more often used in unspecialized articles, while the term “Chatbot” was far more prominent in specialized articles. Other notable differences could be seen in the terms “Algorithm”, “Machine” and “Application”, with the former two more often used in unspecialized articles and the latter more often in specialized articles. “Software” and “System” did not appear substantially more in unspecialized articles than in specialized articles. For the remaining synonyms, there were no clear tendencies distinguishable.

Regarding the expertise of the authors, the preferences were somewhat different, but comparable (figure 8): Non-experts utilized the terms “Algorithm”, “Deep Learning / Neural Networks” and “Machine” significantly more often than experts as well as “Program” and “Software” exclusively, i.e. experts did not use these two terms as synonyms in their articles. In turn, they used “Computer” and “System” far more and “Application” and “Chatbot” noticeably more than

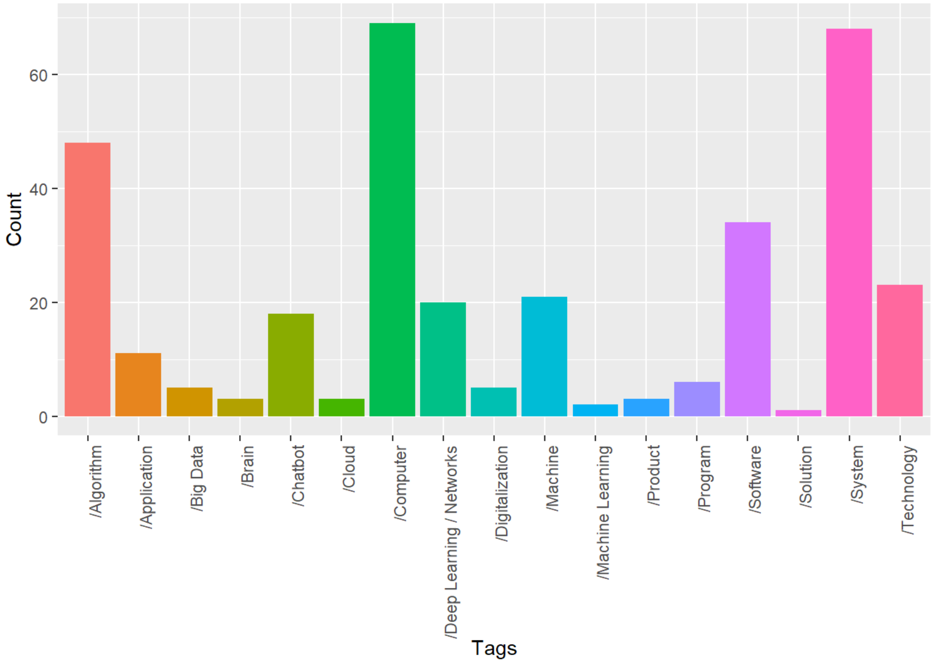


Figure 5: Synonyms

non-experts. The remaining synonyms were again not remarkably distinct in their usage.

For a temporal difference, figure 6 shows the count of the synonyms for every year. Interestingly, the terms “Algorithm”, “Application”, “Computer”, “Deep Learning / Neural Networks”, “Machine” and “Software” were used at least once per year. “Algorithm” was strikingly more often used in articles written in 2019 and 2020 than in earlier articles, with a peak in 2020. Similarly, “Computer” shows also a clear popularity in 2019 and 2020, but instead with an evident peak in 2019. The only clear linear rise in utilization can be seen for “Deep Learning / Neural Networks”, with a more rapid rise in 2020. Comparably, the term “Machine” gained in significance over the years, again with a rapid rise in 2020. In contrast to that, the usage of “Software” was not as linear, with a clear favoritism in 2018 and 2020. A last remarkable observation could be made for the synonym “System”, which was only used in articles written in 2018 and later, with noticeably more mentions in 2019 and 2020.

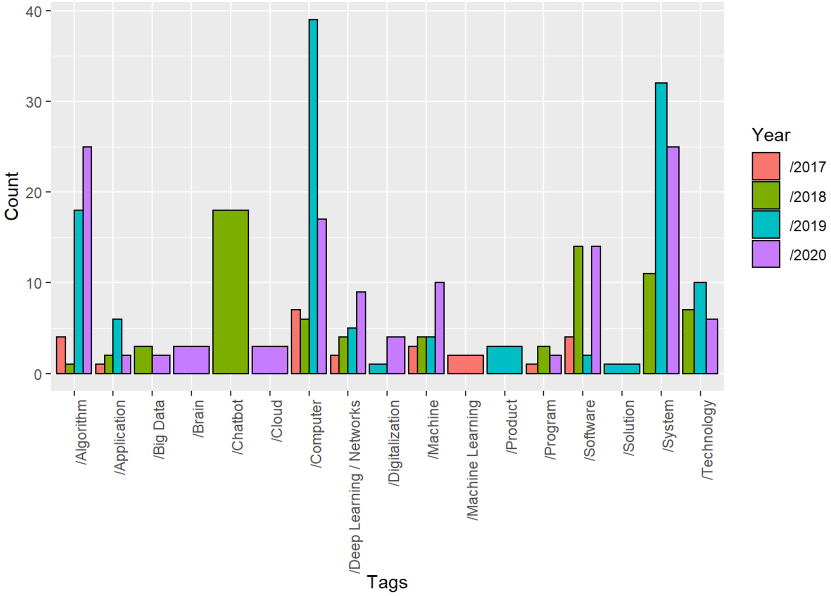


Figure 6: Year and synonyms

5 What to keep in mind

5.1 Limitations

We do not claim to be representative of all articles written in this time period, as our findings were limited in some aspects. To begin with, we had a really small number of articles and they all stemmed from a specific time period. Therefore it would be possible that other articles shed quite a different light on AI in diagnostics, whether they would be from the specific time period or older. As most newspapers did not give free access to their older articles, we could not include them. In the light of the current importance of science and research in terms of reliable information concerning Covid-19, an exhaustive and multi-perspective overview might rather represent the ongoing trend towards a positive sentiment of AI and further research in this field of science.

The distribution of articles in regard to the specialization of the newspapers and the expertise of the authors was also highly unequal, as we had noticeably more articles from unspecialized newspapers and authors. Further, none of the expert authors were AI experts. This was not originally planned, but as none of the articles we found and chose were written by an AI expert, we chose to keep it that way and focus on the medical perspective. AI experts nonetheless might

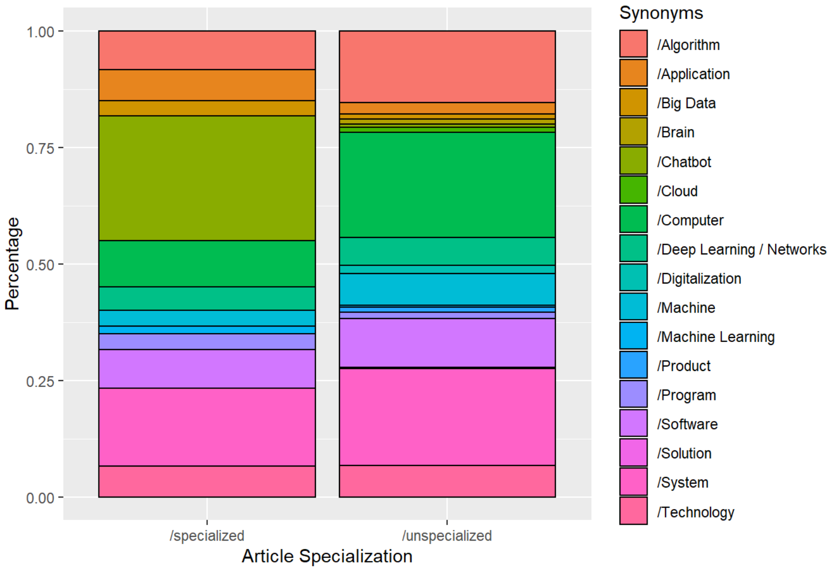


Figure 7: Synonyms and Specialization of Articles

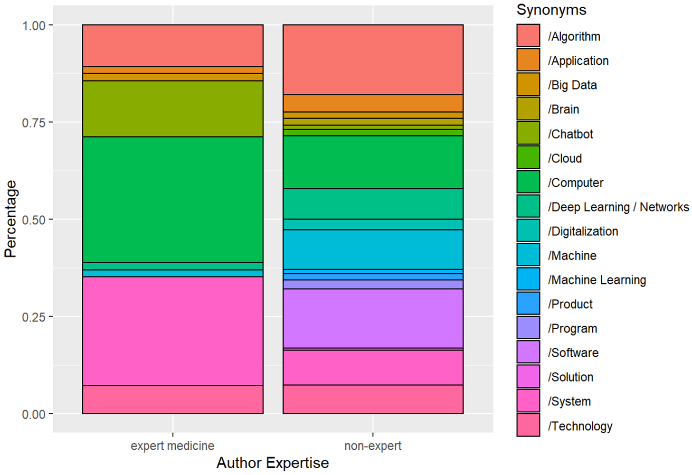


Figure 8: Synonyms and Expertise of Author

would have brought up different opinions and aspects, which we inadvertently excluded. Additionally, some newspapers did not mention the author of the articles, making it harder to understand their perspective and familiarity with the topic and thus limiting the feasibility of analyzing their arguments.

Another limitation was given by the way we searched for the articles. Although we tried not to give too much bias in a specific direction through our search-keywords - which is why we included "neutral" keywords, e.g. "AI in medicine" and more emotionally biased keywords, e.g. "AI replaces physicians" - they of course influenced the articles the search engine showed. Generally, it would be important to mention that a procedure like ours could never be completely free of subjective bias - all of us are involved with AI and its performance to at least some extend through our studies which led us to a rather positive bias towards AI and its applications. This "bias bubble" surely had an effect on the article selection and our manual tagging and therefore distortions of the results would be possible. Even though we tried to counteract it by cross-referencing, we could not rule out the possibility that our manual annotations would vary if we had distributed the articles differently in the first place as we did not specify hard rules on what has to be tagged as what. We would argue here that this subjectivity and lack of clear definitions of boundaries in natural language was a common limitation of manual tagging.

Probably the most important point to keep in mind here would be that our kind of sources did not necessarily represent public discourse accurately and in its entirety. All of the analyzed texts went through the process of preparing information for the publication in press which means that the information displayed in the articles was thought-through content written by people whose profession it is to report news in a concise and adequate way and that the articles were probably not "blindly" published - maybe in contrast to a "simple" tweet by a private person. But it also would mean that the articles might be biased through the author's and newspaper's views and intentions - they intended to catch attention and oftentimes polarize after all.

5.2 Conclusion

In our research we aimed to portray the picture the public draws of AI in context of medical diagnostics and to show which applications are currently most reported on. From our qualitative and quantitative analysis of specialized and unspecialized German news articles we could conclude that not all our expectations were fulfilled - if anything, they were mostly violated. The results indicated a high number of unspecialized articles which brought up surprisingly good arguments. There was no significant difference in sentiment between unspecialized and specialized platforms. Hence our expectations of non-experts being more negative and experts being more positive were not confirmed either.

Most articles tended to report positively about the potential of faster and better diagnostics due to the application of AI. Often highlighted was the reduction of costs due to less needed surgeries and examinations because of it, as well as the opportunity of a more efficient health system. Another important aspect was the location-independent availability of knowledge that AI could provide. A positively reported future application was the production of digital twin organs to help physicians to derive patient-specific properties of an organ.

In contrast to our expectations that non-experts would emphasize the dehumanisation of medicine due to AI, there were multiple articles which stated the contrary - AI could potentially make medicine more humane.

Contra arguments were formulated upon all platforms and oftentimes resulted in a concrete demand of action. The inability of AI to accurately model medical knowledge and experience as well as its lack of transparency and the risk of faulty results were some technical limitations which were brought up. Major concerns were the lack of standardized regulations concerning studies about AI and the potential data misuse, for example by big corporations.

Demands of action were found in 17 out of 24 articles. Most often pointed out was the indispensability of the patient safety as the top priority, as well as the need for legal frameworks, including the open question of who would be legally responsible if the AI made mistakes.

Regarding the *SWOTs*, specialized articles seemed to mention more opportunities but less weaknesses than unspecialized articles, with the former focusing on the shortage of research and the remaining uncertainties surrounding the implementation of AI, whereas the latter concentrating on AI's "technical" properties. In contrast, threats and strengths were equally often communicated. There was an evident development over the years which showed that with every year remarkably less strengths and more threats were written about. In contrast, opportunities remained a consistent topic whereas the amount of referenced weaknesses increased in articles written in the years 2019 and 2020.

As one of our assumptions was to read frequent mentions of AI's dangers, we were surprised to find threats rather infrequently stated. We especially thought that the statement of AI replacing humans in the future would be mentioned regularly, yet our expectation was again violated as this was only brought up by two articles, one of which directly refuted this fear. Threats reported by experts mostly concerned the diagnostic process, while non-experts emphasized legal issues and possible changes in society. Notably, none of the articles written in 2017 communicated any threats at all and those written in 2018 only referred to threats concerning the rise of AI-based health apps, whereas later articles address the blind trust in AI recommendations as the primary threat.

Regardless of this, we identified 17 distinct synonyms rather than explicit descriptions of AI. Certain forms or similar were mentioned, that we wanted to display as they might lead to misconceptions of AI. For instance, "Digitalization" is, if anything, a prerequisite for AI but not AI itself.

Due to the uneven amount of unspecialized and specialized articles, further research is needed to compensate for this for a more reliable outcome. Additionally, as the medical experts and specialized platforms were underrepresented in our sources, we would suggest to include more such articles as well as articles written by AI experts for a potentially different perspective. A shift from our present results could also arise once the current Covid-19 pandemic forfeits its huge impact on humanity, leading to a different portrayal of and sentiment towards AI in medicine. This could also be investigated in future research. Concerning the temporal development in our findings as described above, it would be interesting to investigate how and why they emerged to understand the reasons for a changing attitude towards AI. Finally, a further analysis including several types of media like Twitter or Reddit, additional to newspapers, might convey a more extensive overview on AI in public discourse.

References

Beeger, Britta. 2017. *Diagnosen von Watson.* https://www.faz.net/aktuell/wirtschaft/kuenstliche-intelligenz-soll-krebs-diagnostizieren-15054102.html (31 March, 2021).

Bensch, Hendrik. 2017. *Künstliche Intelligenz eröffnet neue Ansätze in der Psychiatrie.* https://www.bibliomedmanager.de/news/33492-kuenstliche-intelligenz-eroeffnet-neue-ansaetze-in-der-psychiatrie (31 March, 2021).

Buck, Christian. 2019. *Dr. Algorithmus.* https://www.welt.de/wirtschaft/bilanz/online189418949/Dr-Algorithmus-Kuenstliche-Intelligenz-in-der-Medizin.html (31 March, 2021).

Cambridge Dictionary. 2021. *Digital Divide.* https://dictionary.cambridge.org/de/worterbuch/englisch/digital-divide (31 March, 2021).

Focus Online. 2019. *Künstliche Intelligenz kann Ärzten bei Analyse von Röntgenbildern helfen.* https://www.focus.de/digital/dldaily/digital-health/16-000-roentgenbilder-bewertet-kuenstliche-intelligenz-hilft-aerzten-bei-analyse-von-roentgenbildern_id_10216957.html (31 March, 2021).

Forschungszentrum Jülich. 2019. *Künstliche Intelligenz hilft, Schizophrenie besser zu verstehen.* https://www.fz-juelich.de/SharedDocs/Pressemitteilungen/UK/DE/2019/2019-12-03-ki-hilft-schizophrenie-besser-zu-verstehen.html#:~:text=J%C3%BClicher%20Hirnforscher%20haben%20nun%20mithilfe,den%20einzelnen%20Patienten%20zugeschnitten%20ist (31 March, 2021).

Healthcare in Europe. 2020. *Studie zeigt großes Potenzial von KI-Kooperation.* https://healthcare-in-europe.com/de/news/studie-zeigt-grosses-potenzial-von-ki-kooperation.html (31 March, 2021).

Herbe, Ann-Christin. 2018. *KI in der Medizin: Der Computer weiß, was dir fehlt.* https://www.dw.com/de/ki-in-der-medizin-der-computer-wei%C3%9F-was-dir-fehlt/a-46210336 (31 March, 2021).

Kaulen, Hildegard. 2020. *Überschätzte KI: Sind Algorithmen tatsächlich die besseren Ärzte?* https://www.faz.net/aktuell/wissen/medizin-ernaehrung/ueberschaetzte-ki-sind-algorithmen-tatsaechlich-die-besseren-mediziner-16754548.html (31 March, 2021).

Klöckner, Jürgen. 2019. *TK-Manager: Deutschland steht kurz vor einer Medizin-Revolution.* https://www.focus.de/digital/dldaily/digital-health/interview-zur-digitalisierung-tk-manager-ueber-medizin-revolution-diagnosen-werden-nicht-mehr-ohne-ki-getroffen_id_10592305.html (31 March, 2021).

Kröplin, Tim. 2018. *Dr. KI hat nun Zeit für Sie.* https://www.zeit.de/wissen/gesundheit/2018-11/bilderkennung-kuenstliche-intelligenz-gesundheit-arzt-diagnose-smart-devices? (31 March, 2021).

Kuhn, Sebastian, Stefanie Maria Jungmann & Florian Jungmann. 2018. *Künstliche Intelligenz für Ärzte und Patienten: „Googeln“ war gestern.* https://www.aerzteblatt.de/archiv/198854/Kuenstliche-Intelligenz-fuer-Aerzte-und-Patienten-Googeln-war-gestern (31 March, 2021).

Matera, Elena. 2020. *Wie Künstliche Intelligenz den Krebs besiegen soll.* https://www.berliner-zeitung.de/zukunft-technologie/wie-kuenstliche-intelligenz-den-krebs-besiegen-soll-li.118424 (31 March, 2021).

Rabhansl, Christian & Christian Maté. 2020. *Künstliche Intelligenz und Medizin - „Es kommt darauf an, dass Arzt und Maschine ein gutes Team bilden“.* https://www.deutschlandfunkkultur.de/kuenstliche-intelligenz-und-medizin-es-kommt-darauf-an-dass.1270.de.html?dram:article_id=474497 (31 March, 2021).

Redaktionsnetzwerk Deutschland. 2020. *Künstliche Intelligenz im Krankenhaus: Wie sie die Diagnostik verändern kann.* https://www.rnd.de/gesundheit/kunstliche-intelligenz-im-krankenhaus-verandert-die-krebsmedizin-WS2TAJXLV5AUNOA4FZADVWTEUI.html (31 March, 2021).

rme/aerzteblatt.de. 2017. *Arzt versus Computer: Wer erkennt Brustkrebsmetastasen am besten?* https://www.aerzteblatt.de/nachrichten/87011/Mi-Arzt-versus-Computer-Wer-erkennt-Brustkrebsmetastasen-am-besten (31 March, 2021).

Röhrlich, Dagmar. 2020. *Algorithmen in der Medizin - Wenn Computer besser diagnostizieren als Ärzte.* https://www.deutschlandfunk.de/algorithmen-in-der-medizin-wenn-computer-besser.724.de.html?dram:article_id=472398 (31 March, 2021).

Stratmann, Gerrit. 2020. *Künstliche Intelligenz in der Medizin: Kann KI den Krebs besiegen?* https://www.deutschlandfunkkultur.de/kuenstliche-intelligenz-in-der-medizin-kann-ki-den-krebs.976.de.html?dram:article_id=480641 (31 March, 2021).

strz/Esanum. 2018. *Prostatakrebs: Ultraschall und Künstliche Intelligenz liefern Spitzenergebnisse bei Diagnostik.* https://www.esanum.de/today/posts/prostatakrebs-ultraschall-und-kuenstliche-intelligenz-liefern-spitzenergebnisse-bei-diagnostik (31 March, 2021).

Till, Ulrike. 2020. *Künstliche Intelligenz im Kampf gegen Krebs.* https://www.swr.de/wissen/ki-krebs-100.html (31 March, 2021).

Wikipedia Authors. 2021a. *Elektronische Gesundheitsakte.* https://de.wikipedia.org/wiki/Elektronische_Gesundheitsakte# (31 March, 2021).

Wikipedia Authors. 2021b. *Watson.* https://en.wikipedia.org/wiki/Watson_(computer)#Healthcare (31 March, 2021).

Witte, Felicitas. 2019a. *Dr. Algorithmus: Wie künstliche Intelligenz Krebspatienten helfen kann.* https://www.focus.de/digital/dldaily/digital-health/hoffnung-fuer-kranke-dr-algorithmus-wie-kuenstliche-intelligenz-krebspatienten-helfen-kann_id_10467456.html (31 March, 2021).

Witte, Felicitas. 2019b. *Ein Arzt fordert: Wir müssen Computer mit den Wünschen der Patienten füttern.* https://www.focus.de/digital/dldaily/digital-health/kuenstliche-intelligenz-in-der-medizin-den-computer-mit-den-wuenschen-der-patienten-fuettern_id_10514832.html (31 March, 2021).

Witte, Felicitas. 2019c. *Im Kampf gegen Krebs sind wir "mitten in einer explosiven Entwicklung".* https://www.focus.de/digital/dldaily/digital-health/onkologie-wir-sind-mitten-in-einer-explosiven-entwicklung_id_10570700.html (31 March, 2021).

ZEIT online. 2019. *Wenn Computer Röntgenbilder auswerten.* https://www.zeit.de/wissen/2019-09/kuenstliche-intelligenz-medizin-diagnose-krankheiten-bilddiagnostik (31 March, 2021).

Chapter 18

AI in healthcare – Expectation vs. reality of breast cancer detection

Tim Bax, Milan Ewert, Florian Pätzold & Franka Timm

Medical complexity increased over the years, ever higher standards in the healthcare system put science under pressure to meet the requirements and to provide adequate and qualitative high care for patients. The application of Artificial Intelligence (AI) happens to be a promising tool in order to improve accuracy and efficiency in medicine. This new upcoming scientific field affects the personal space of many people in their everyday lives. There is an interesting crosspoint of scientific reliable knowledge and expectations people set for themselves or grounded on supposedly trusted sources like news articles. Exactly this phenomenon could be observed in 2020 when AI was applied to breast cancer detection for the first time. In the following, it will be discussed how expectations and reality differ in public discourse concerning AI and its application for breast cancer detection within the healthcare system. It will be evaluated what similarities and differences are shown between several public sources and scientific articles. Our results show that online articles published by news outlets and blogs mainly correspond to the key aspects that the original paper about the AI system provides.

Keywords: Artificial Intelligence | Healthcare | Breast Cancer Detection | Google Health | Expectations

1 Introduction

In the year 2020 there has been revolutionary progress in digital mammography due to the implementation of an AI system by McKinney et al. 2020 that provides automated computer-aided detection of breast cancer. This technology paves the way for clinical trials to improve the accuracy and efficiency of breast cancer screening. With the aim to reduce the workload of radiologists and decrease false diagnoses AI should enhance quality in the healthcare system as well as improve patient care. Nevertheless, the application of such an AI causes a lot of unanswered questions and expectations including fears and hopes expressed by the public. In many cases these expectations towards AI seem to surpass reality. Even though it is difficult to predict to what extent AI will shape the future of the scientific field of breast cancer detection, in the following we will discuss how expectations expressed in online news and blog articles, representative for the general public, correspond with the reality towards AI regarding the current state of the art of breast cancer detection. More specifically, it will be shown how the use of AI is represented in public discourse, for instance in news articles, in comparison to the scientific reality of AI systems in breast cancer screening.

2 Methodology

In order to compare the above mentioned expectations against the reality of the original paper we investigated 15 more wide-spread and well established news articles as well as 14 smaller, more subjective healthcare and technology-related blog and magazine articles. The authors of such articles operate as journalists as well as private agents, some of which show middle to high professional expertise in the area of technology and healthcare. The articles are selected by searching online exclusively with the keywords ”google”, ”ai”, ”breast”, ”cancer”, ”detection”, using the search engines Google and DuckDuckGo. Afterwards, we chose top result articles that appeared on top of the screen without limiting to a specific country, but excluding those bound to a subscription in order to be read. The news and blog articles represent the public expectation on the possible use of AI in breast cancer detection, which is then compared to the actual strengths and potentials the original paper ”International evaluation of an AI system for breast cancer screening” (McKinney et al. 2020) provides. Subsequently, our findings are discussed and interpreted in a qualitative way by using a method that makes it more convenient to compare the articles: we defined a tag-set for the content, including the tags ”strength”, ”weakness”, ”opportunity”, ”demand”, ”threat” (SWODT) and analyzed all sources qualitatively by tagging the statements in the respective online articles using the ”Computer Assisted Text Markup and Analysis” (CATMA) (Meister et al. 2019) online tool for textual annotations. Afterwards, we used Excel to generate plots that visualize and underline our findings for better comprehensibility.

3 Findings

3.1 Expectations

Started in 2006, Google Health was originally designed to create a repository of health records to connect healthcare teams (Wikipedia authors 2021). Today, Google's branch is developing technology solutions for enhancing quality of patient care. These include artificial intelligence for breast cancer screening and major news articles now emphasize the strengths as well as the opportunities such a system may offer. According to all reviewed articles, by applying the DeepMind AI algorithm the study has shown that the reduction of false positive (5.7 and 1.2 percent) and false negative diagnoses (9.4 and 2.7 percent) in the US and the UK respectively, prove the superiority of the AI system compared to a single radiologist reading the mammography scans. Also, the authors go on, as AI does not get tired and has greater computing power than a human being, even in a thousand cases the system is able to stay focused and analyze every pixel of a scan that humans are not even able to perceive (Park 2020, Thomson Reuters 2020, Griffin 2020). Even though no patients' history or prior mammograms were used, the algorithm exceeded human performance. Already today, there are substantial advantages over human performance that are underlined by the news articles; still further opportunities are addressed, the AI may be able to fulfill. For example, if additionally the system is pre-loaded with patient history, the progress could even be further approved (Reid 2020). Moreover, Walsh 2020, Reid 2020 and Lovett 2020 highlight the possibility that a great reduction of workload could be achieved by substituting the dual reading process currently being commonly practiced in the UK, where two doctors each read the mammograms at hand instead of just one. As humans still could not be replaced, Collins 2020, Thorbecke 2020, Eddy 2020 and Samuel 2020 propose a workframe in which Google's AI and professional radiologists work closely together to improve the healthcare system in accuracy of reading breast cancer scans as well as in a decrease of wait times and patient stress.

Although these claims seem very promising, all articles also mention that the AI still has to be approved and more research is needed to improve patient care. This could take up several years. It is heavily demanded to look at the system as a general support rather than a standalone instance in the healthcare system and Ellis 2020 for instance suggests that this process should not be pushed inconsiderately. Following this request, some articles also raise their concerns regarding some weaknesses and even threats of the DeepMind algorithm. Such weaknesses include the fact that the study only used limited data from two hospitals in the

UK and one hospital in the US, questioning the generalizability of the AI for other demographic parties (Ellis 2020). Furthermore, as the neural networks of such an algorithm are trained on large datasets, privacy issues may arise from learning when processing the patients' information (Ellis 2020, Manskar 2020). Also addressed is the problem that such systems often represent a so-called black box that not even the developers seem to understand in detail. It would therefore be difficult to apply this system to such complex decisions as the detection and classification of breast cancer, which often need to be carefully weighed in order to arrive at a proper diagnostic analysis. Otherwise, by diagnosing potentially harmless ulcers, anxiety can be caused for affected patients. and to not unnecessarily cause anxiety for patients by diagnosing potentially harmless ulcer. The generalizability for other countries is furthermore challenged by the fact that the research team exclusively investigated cases where the same imaging equipment was used (Thomson Reuters 2020, Griffin 2020). Other techniques may result in different, possibly false diagnoses made by Google Health's algorithm. There is also a difference between retrospective studies where the patient's final diagnosis is already known - as in this breast cancer study by McKinney et al. 2020 - and getting the AI to make exact classifications of current patients whose diagnoses are yet unknown, which is the case for the prospective diagnosis process. To evaluate how the system works in the real world, further prospective studies have to be conducted (Samuel 2020).

Many blog articles include the majority of the already mentioned aspects and are therefore mostly consistent with the news. Nevertheless, some of them contain points that have not been mentioned yet. One point of criticism is that doctors may rely too heavily on the diagnosis of the AI (Ray 2020). The machine is only calculating a continuous probability value instead of making a binary judgment whereas doctors have to decide which next step will be the best for the patient. (Ray 2020, Vincent 2020). Even if the probability for cancer might be low and the AI system would not diagnose cancer, some patients might appreciate the prescription of a biopsy to ensure a correct diagnosis (Ray 2020). Another mentioned aspect is that the AI system, even if it acts as a safety net and widely helps radiologists with their diagnosis, will not solve the shortage of qualified personnel (Sarkar 2020). While the system could help to improve the probability of finding cancer it could also lead to over-diagnosis of cancer (Vincent 2020). Once the diagnosis ”cancer” is made, many costly, painful and potentially life-changing medical interventions could follow that would not be necessary if the small ulcer found would not harm people over their lifetime anyway (Vincent 2020). That is because the complexity of the real world is overlooked (Machemer 2020, Vincent 2020). Even if the system spotted cases of cancer that doctors did not spot, it also missed cases that were not missed by doctors (Ray 2020, Brodwin 2020). Furthermore, Bisen 2020 commented that the demographics of the population studied by the research team are not well defined in the previous AI-based detection. The performance of AI algorithms can be highly dependent on the population used in the training data sets which could lead to a racial bias of the algorithm (Machemer 2020). Also, the cancers classified by the AI more often required invasive care compared to those identified by the radiologist for which the researchers did not have an explanation (Board 2020).

3.2 Reality

In order to compare the content of the article discussed above to which we refer to as 'expectations' we want to find something that is the scientific 'reality'. The article "International evaluation of an AI system for breast cancer screening" by McKinney et al. 2020 is concerned with a study that evaluates the performance of a new AI system for breast cancer prediction using two large, clinically representative datasets from the UK and the USA in order to compare the predictions of this system to the ones made by readers in routine clinical practice. Furthermore, the study wants to show that the performance of this AI system is able to exceed the performance of individual radiologists. But how is this accomplished in reality? An AI system was trained to identify the presence of breast cancer from a set of screening mammograms and was evaluated in three ways. First, AI predictions were compared with the historical decisions made in clinical practice. Second, to evaluate the generalizability across populations, a version of the AI system was developed using only the UK data and retested on the US data. Finally, the performance of the AI system was compared to that of six independent radiologists using a subset of the US test set.

Right in the beginning, the article predicts that this AI system can easily outperform a human radiologist if there is enough training data. This might increase the likelihood of the assumption that the application of AI in the medical healthcare system might replace doctors someday. However, in the later course of the article, Scott Mayer McKinney et al. broached the issue of potential clinical applications and performance breakdowns of the AI system. They conclude that the AI system could be used to reduce the workload. In the double-reading process, a procedure practiced in the UK where two individual radiologists analyse a single mammogram, this would mean that there is a possibility to omit the second reader when the decision of the AI system is in-line with the decision of the first reader. Moreover, The AI system could also be used to provide automated, immediate feedback in the screening setting. All in all, comparing the errors of the AI system with errors from clinical readers revealed many cases in which the AI system correctly identified cancer whereas the reader did not, and vice versa. Investigating the results, due to breakdowns of the system when it comes to classifying different cancer types, it has become clear that the AI system is very sensitive especially for the identification of invasive cancers rather than in situ cancers. In situ cancer describes a group of abnormal cells that are not classified as cancer yet but have high potential to develop cancer later. In contrast invasive cancer is cancer that has spread already to tissues beyond the tissue where it developed originally, hence growing into surrounding, former healthy tissues.

In the discussion, the article explains that the AI system could not outperform the accuracy of diagnoses in the case of double-reading but it was statistically non-inferior to the performance of the second-reader. Furthermore, they emphasize that on the one hand it has to be noted that in comparison to the human readers the AI system has no access to the patient history or previous mammograms when making a screening decision. On the other hand, all of the radiologists who were chosen for screening mammograms did not uniformly receive fellowship training in breast screening. As a consequence, it has to be considered that in the case of more specialized human readers the benchmark could have been much higher. To conclude, the article claims that there will be a need for more clinical studies in order to understand the full extent to which this technology can benefit patient care. For example they mention that a promising tool for further medical application would be local fine tuning in order to accomplish a stronger baseline performance of the system (McKinney et al. 2020).

4 Analysis and Discussion

The major news articles show great variety in detail regarding the content of AI in breast cancer screening. Undeniably, most of the articles show a general tendency to emphasize not only the strengths but also the great opportunities of Google's AI system for early breast cancer detection.

The above mentioned SWODT-analysis proved the overall positive public receivement of the paper. In both news and blog articles combined, possible strengths and opportunities of the AI were mentioned 147 times, compared to only 64 weakness and threat statements (see Figure 1). Strikingly, despite the fact that almost all authors agree that the AI system is far from being ready to be introduced to the healthcare system, let alone multiple different healthcare systems, the demands to the AI system are more of a side note with only 33 mentions, mostly with just one instance per article. Comparing the two subgroups - news and blog articles - in the SWODT-analysis, one can observe a pronounced difference in the appearance of demands and weaknesses. While the news articles elaborate on the demands (news: 23, blogs: 10) and opportunities (news: 38, blogs: 25) way more thoroughly, the opposite is true for the weaknesses (news: 14, blogs: 31) (see Figure 2). However, since all the other subjects of the SWODT-analysis are comparable, we denounce the severe differences between demands and weaknesses to two outliers, namely Vincent 2020 and Ray 2020, which alone mentioned 13 weaknesses and no demands.

Furthermore, the popular topic of data security when it comes to any technology news was non-apparent. Most articles neglected possible threats like data security or over-diagnosis problems entirely, as this was the least talked-about topic regarding the AI in the present sources. Schroeder 2020 was the only out of 29 articles that reported about "Deep Mind [allegedly being] granted access to the healthcare data of 1.6 million patients in the UK's National Health System (NHS) without explicit permission". Another notable point is that most articles did not differentiate between one or two doctors. While it is true that the AI is able to exceed the performance of a single radiologist, this is not the case for a double reading system as it is the case in the UK. Here the performance was as good as the doctors but not better. While the word ”better” in comparison to doctors performance was used 24 times, especially in attention-seeking headlines and the beginnings of the articles, it was clarified only 15 times that this statement is referring to an individual radiologist and is wrong otherwise. The fact that information is clearly communicated in the paper of McKinney et. al, indicates that the public discourse of AI in breast cancer detection is generally well summarized but in some cases important context details are left out by public online articles.

Where expectations and reality differ, misleading deceptions could form a precarious false picture of AI in the general public. Nevertheless, as the main important points of the paper are mentioned by most of the articles and all articles presented the correct data of the study, giving the reader the possibility to form its own unbiased opinion, such a false picture does not develop for the case at hand.

5 Conclusion

In this short article we investigated the difference between expectations and the reality of Google Health's AI system for breast cancer detection in the healthcare

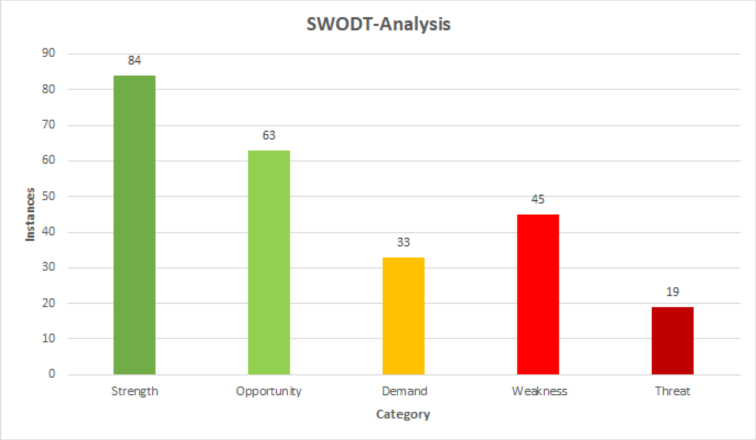


Figure 1: The "SWODT" analysis of all articles.

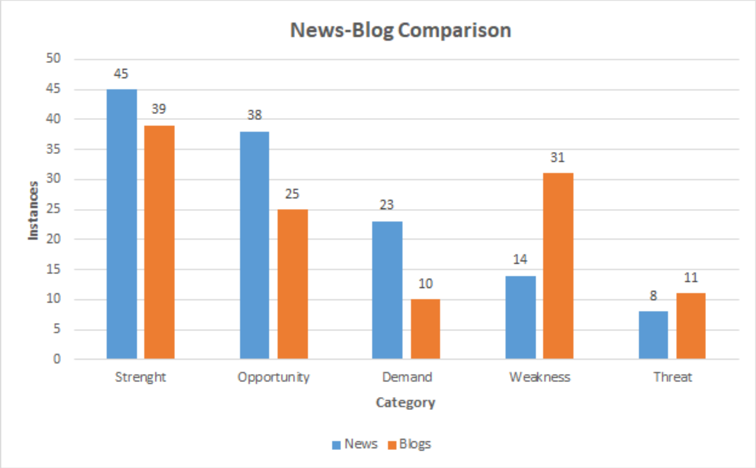


Figure 2: Comparison of news and blog articles.

system. By conducting a "SWODT" tagset analysis we were able to derive several interesting results.

Most articles focus on the possible strengths and opportunities the AI could bring to the healthcare system. However, there is a higher tendency for instances of strength, opportunity and demand in the news articles whereas the blog articles are more critical. Strikingly, there are only two cases, both being blog articles, where the negative content outweighs the positive, thus leading to distortion in the news and blog article comparison. All papers that elaborate on the implementation into the healthcare system agree that the AI should not replace doctors but be used in tandem with human radiologists. Also it is highlighted that the system is far from being ready to be implemented into the healthcare system and that more research and studies are needed, which will take several years. These main points are congruent with the online news and blogs, such that in consideration of all articles, one can say that expectations and reality are reasonably similar to not form a false picture of healthcare's potential AI systems, in the public.

References

Bisen, Vikram Singh. 2020. *How does Google AI detect breast cancer better than radiologists?* https://www.vsinghbisen.com/technology/ai/how-google-ai-detect-breast-cancer/.

Board, Advisory. 2020. *Google's AI beats doctors at detecting breast cancer. (Except when it doesn't.)* https://www.advisory.com/daily-briefing/2020/01/06/google-ai.

Brodwin, Eric. 2020. *Google AI tool can pinpoint breast cancer better than clinicians.* https://www.scientificamerican.com/article/google-ai-tool-can-pinpoint-breast-cancer-better-than-clinicians/.

Collins, Katie. 2020. *Google health's AI can spot breast cancer missed by human eyes.* https://www.cnet.com/news/google-healths-ai-can-spot-breast-cancer-missed-by-human-eyes/.

Eddy, Nathan. 2020. *Google AI platform aids oncologists in breast cancer screenings.* https://www.healthcareitnews.com/news/google-ai-platform-aids-oncologists-breast-cancer-screenings.

Ellis, Rachel. 2020. *As artificial intelligence proves as effective as doctors at reading scans, we ask... would you trust a computer to diagnose your illness?* https://www.dailymail.co.uk/health/article-8067601/Would-trust-computer-diagnose-illness.html.

Griffin, Andrew. 2020. *Google AI can spot possible breast cancer better than trained experts and could dramatically improve detection, study suggests.* https://www.independent.co.uk/life-style/gadgets-and-tech/news/google-ai-breast-cancer-artificial-intelligence-deepmind-healthcare-a9267081.html.

Lovett, Laura. 2020. *Study: Google's AI tech shows promise in detecting breast cancer.* https://www.mobihealthnews.com/news/study-googles-ai-tech-shows-promise-detecting-breast-cancer.

Machemer, Theresa. 2020. *What does Google's breast cancer screening A.I. actually do?* https://www.smithsonianmag.com/smart-news/what-does-googles-breast-cancer-screening-ai-actually-do-180973907/.

Manskar, Noah. 2020. *Google AI spots breast cancer better than doctors, study finds.* https://nypost.com/2020/01/02/google-ai-spots-breast-cancer-better-than-doctors-study-finds/.

McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577(7788). 89–94.

Meister, Jan Christoph, Marco Petris, Christian Bruck, Malte Meister, Marie Flüh, Jan Horstmann, Janina Jacke, Mareike Schumacher & Evelyn Gius. 2019. *CATMA*. Version 6.0.0. DOI: 10.5281/zenodo.3523228. https://doi.org/10.5281/zenodo.3523228.

Park, Alice. 2020. *Google's AI bested doctors in detecting breast cancer in mammograms.* https://time.com/5754183/google-ai-mammograms-breast-cancer/.

Ray, Tiernan. 2020. *Is Google breast cancer detection AI better than doctors? Not so fast.* https://www.zdnet.com/article/google-deepminds-ai-based-breast-cancer-detection-is-not-yet-an-automatic-diagnostician/.

Reid, David. 2020. *Google's DeepMind A.I. beats doctors in breast cancer screening trial.* https://www.cnbc.com/2020/01/02/googles-deepmind-ai-beats-doctors-in-breast-cancer-screening-trial.html.

Samuel, Sigal. 2020. *AI can now outperform doctors at detecting breast cancer. Here's why it won't replace them.* https://www.vox.com/future-perfect/2020/1/3/21046574/ai-google-breast-cancer-mammogram-deepmind.

Sarkar, Prapti. 2020. *Google's AI outperforms radiologists at detecting breast cancer.* https://www.shethepeople.tv/news/google-artificial-intelligence-breast-cancer/.

Schroeder, Stan. 2020. *Google's AI is better at breast cancer screening than human experts, study claims.* https://mashable.com/article/google-ai-breast-cancer/?europe=true.

Thomson Reuters. 2020. *AI detects breast cancer as accurately as expert radiologists, study finds.* https://www.cbc.ca/news/health/ai-mammograms-breast-cancer-1.5412679.

Thorbecke, Catherine. 2020. *AI-powered computer 'outperformed' humans spotting breast cancer in mammograms: Study.* https://abcnews.go.com/Technology/ai-powered-computer-outperformed-humans-spotting-breast-cancer/story?id=68031128.

Vincent, James. 2020. *Why cancer-spotting AI needs to be handled with care - Accelerating cancer diagnoses could hurt more than it helps.* https://www.theverge.com/2020/1/27/21080253/ai-cancer-diagnosis-dangers-mammography-google-paper-accuracy.

Walsh, Fergus. 2020. *AI 'outperforms' doctors diagnosing breast cancer.* https://www.bbc.com/news/health-50857759.

Wikipedia authors. 2021. *Google Health - Wikipedia.* https://en.wikipedia.org/wiki/Google_Health.

Chapter 19

Should robots take care of the elderly? – Comparing ethical guidelines to real life experiences

Rabea Breininger, Luisa Drescher, Lilith Okonnek & Inga Wohlert

Due to the care crisis in Germany and the ageing society, elderly care has been a highly relevant topic in public discourse. As an approach for a solution care robots became part of this discussion. In this chapter, we will present a qualitative analysis of the discourse in documentaries compared to a statement by the German Ethics Council about care robots. To collect labelled data, a modified SWOT-Tagset was used. The results show that there are both similarities and differences between the documentaries and the Ethics Council. The topic is most often approached in a differentiated manner and from opposing perspectives.

Keywords: AI | Public Discourse | Elderly Care | Robotics | German Ethics Council

1 Introduction

Many fields of nursing care are lacking qualified employees. According to the German Federal Ministry for Health, the number of people in need of care will be approximately 6.1 million by the year 2050 (Bundesministerium für Gesundheit 2021). The reason for this is the demographic change and the associated increase in life expectancy. Additionally, nursing care is not a very attractive job for many people and only a small number of school leavers actually take a career as a nurse into account. The working conditions are often associated with extra hours, night shifts, low salary, but also with heavy lifting and physical health problems arising therefrom. Therefore, there is a significant lack of staff and this lack is predicted to increase even further. Studies have shown that there are only 11 full time nurses for every 100 over 80-year-olds in Germany (OECD 2011). These developments represent significant challenges and require adjustments and a lot of work.

One approach to address this problem is to make use of artificial intelligence (AI) to help with caring. There is not a single definition of what AI is. In general, the term AI refers to any human-like intelligence exhibited by a computer, robot or other machines. There are two kinds of AI that are commonly differentiated, namely weak and strong AI. Strong AI aims to develop machines that have similar abilities as humans with respect to self-aware consciousness, the ability to solve problems, to learn and to plan. Weak AI, on the other hand, aims to find solutions for specific problems and relies on human interference to define the parameters of its learning algorithms. Therefore, weak AI rather simulates human-like consciousness and intelligence and is the main approach used for AI research (IBM Cloud Education 2020). In the following, the term ”robot” can be seen as an extension of the concept of AI as robots provide a physical interface of AI and represent its embodiment in the tasks they are designed to solve.

So far, the use of AI in elderly care is an exception and therefore there are only a few pilot studies at hand. For example, the robot Pepper is used in some care facilities primarily to entertain the elderly and to make them familiar with the use of AI. Pepper is able to talk with the care recipient, sing, make jokes, dance and has many further abilities. One of the most valued features of Pepper is that the robot is able to read and react to the people’s emotions. Another well-known research project is a robot that looks like a seal and is called Paro. Paro is designed to act like a living seal baby and to learn during the interaction, especially how the humans behave such that they can adapt to the care recipients. This learning process is achieved with the help of AI learning algorithms. The German Federal Research Ministry is convinced that AI can contribute towards easing the challenging situation in nursing facilities. The German Ethics Council also commented on that and promotes the use of AI in care facilities while proposing ethical guidelines and certain demands to get the most positive result. All in all, there already are a few prototypes used in some facilities, but this is far from being the standard (Birthe Sönnichsen, ARD-Hauptstadtstudio 2020).

The present study investigates which strengths, weaknesses, opportunities, threats and demands are presented by the German Ethics Council as well as by a collection of German documentaries and if they correspond to each other. There will be a closer look at how these sources present their key points and especially if they are discussed in a sophisticated manner. Furthermore, it will be investigated if both sources describe the current status of the use of AI in elderly care in the same way and how they use the term AI.

We predict a different presentation of the use of AI in elderly care presented in the German Ethics Council and the collection of documentaries. Furthermore, we hypothesise that the documentaries present their key aspects in an opinion-forming manner.

2 Methods and Materials

The decision to focus on German sources was influenced by personal connections to the German care system and the previous exposure to the wide discussion of the care crisis in nursing environments at hand.

As the representative of a rational display of opinions on robots in elderly care the statement ‘Robotik für gute Pflege‘ (Deutscher Ethikrat 2020) by the German Ethics Council, published on March 10th 2020, was selected. From the council’s role, defined in the official law ‘Gesetz zur Einrichtung des deutschen Ethikrats (EthRG 2007)‘, an unbiased and rational evaluation of the topic was expected (Bundesamt für Justiz 2007). In this law it is stated that the council is supposed to answer medical, scientific and ethical questions with the consequences for individuals in mind. Suggestions for political reactions and the information of the public pose a great part of their responsibility (Bundesamt für Justiz 2007). Hence, the Ethics Council acts as a suitable agent to analyse the public discourse. As the statement does not explicitly focus on the care for elderly in particular, all arguments that were directed at the entirety of the care systems as well as specific aspects in elderly care were considered. The Ethics Council is funded by the government and can be considered a legislative governance power (Bundesamt für Justiz 2007). Even though their expertise is not necessarily founded in elderly care or Artificial Intelligence, the Ethics Council can be regarded as a professional agent when sharing their statement with the public. Within the statement, it is acknowledged that robotic applications are not equivalent to Artificial Intelligence, but in their recommendations and opinions the statement does not differentiate between the possible forms of robots, therefore they were incorporated alike.

In contrast to the official position of the Ethics Council’s statement, the people directly affected by robots in elderly care are represented by subjective documentaries that allow for an emotional perspective. Multiple videos, accessible online via YouTube, were facilitated to be able to accumulate real shared experiences of developers, caretakers, care recipients and dependants. In addition, the presentation and entertainment purposes of this source, seen in the editing, chosen title and the reporter's personal remarks, give an insight into the contribution to the public opinion. Depending on direct or indirect speech of those affected and the influence of reporters, the agents can vary. They differ from private people who use AI to professionals in AI as developers to members of the media who follow an individual agenda for entertainment. The showcased use of technology includes many applications with different forms of AI, ranging from robots functioning as companions for the elderly, devices for monitoring, entertainers and robots executing heavy lifting to reminders for drinking water and taking medications. The robots, mostly up to the standards for a trial run, were operated and adjusted by researchers working on their development.

Additionally, videos that wrongfully used the label AI were included in the analysis, as this misuse holds further information about the public discourse. Ten short documentaries and two videos longer than ten minutes, published by public service broadcasters and private media outlets, were analysed. Eleven out of twelve sources were broadcasted between 2017 and 2020, one in 2011. The selected videos show interactions between the systems and elderly in different settings. While some portray the function of incorporated technologies or the demonstration in laboratories, many present the viewer with first contacts of elderly or caretakers and robots to capture the first impressions of the interactions.

To capture the nature of the discourse presented in the two different kinds of source materials, a qualitative method was employed. According to the SWOT-Tagset, opinions and statements were sorted into multiple categories and then analysed in their affiliation. The differentiation between strengths, weaknesses, opportunities and threats was further refined, to maximise the information gain, and completed by the tag demand. The latter tag indicates the requirements that have to be met in the sources eyes. The demand label was introduced due to the way especially the Ethics Council presented their suggestions. Using CATMA on the council's statement and noting proclamations by their time in the documentaries, the data was accumulated in association with their complementary tag (Meister et al. 2019). With a qualitative analysis the tag-statement pairs were compared to one another and presented to emphasise the sources' attitude towards robots in elderly care.

3 Results

To get an impression of care robots, the German documentaries shed light on the current situation of AI use in retirement homes and the possibility of using robots in retirement and private homes of elderly people as helpers in an autonomous life. Most of the locations shown were current pilot projects in Germany.

The robots presented were mainly Paro (Tansek 2020), Pepper (Regional 2017), CASERO, Care-O-bot (IPA320? 2011) and GARMI (BR Fernsehen 2019), of which Pepper appeared most often.

Before going into the detailed analysis, we will briefly describe some of the reactions towards the robots hereafter. Taking the robot Pepper as an example, the first impression people have of Pepper can already differ a lot, especially since sometimes the human-like aspects of Pepper seem to be more in focus while at other times the robot features seem to be more essential for the first impression (WDR 2020) (Main-Post 2019). One woman even mentioned that "Pepper" reminded her of her children when they were little (Kroth 2018: 31:19 min). However, there seems to be a difference between the first encounter with Pepper and subsequent encounters. Groups new to Pepper were less accustomed to it but in contrast people who had known Pepper longer saw it more like another person. Nonetheless, it was noted that for some activities, such as discussions, Pepper is not an improvement but rather a setback compared to humans (Plahl 2020: 2:47-3:04 min). The machines designed to help the caregivers with more housekeeping tasks received fewer reactions from the residents of the nursing home. This was due to lesser interactions between residents and robots. Rather, there were more reactions from the staff who responded positively to them. Most of the staff noted that it can spare them from a lot of work and they can spend that time with the elderly (IPA320? 2011). This is just an overview of some of the reactions and sheds little light on the actual discourse. In the following, we will analyse the sources using the SWOT Analysis.

3.1 Tagset Analysis

3.1.1 Strengths

When having a look at the strengths emphasised in both the documentaries and the statement of the German Ethics Council, the support which robots provide to caretakers is an often recurring theme. Here, especially physical support plays an important role, as robots take over strenuous tasks like transporting heavy objects such as beverage crates and laundry (IPA320? 2011: 00:23) or serve as lifting aid (Deutscher Ethikrat 2020: p.16). Apart from exhausting tasks, robots can take on multiple smaller duties that are part of everyday work. For instance, there are computer systems that are able to take out the trash and deliver medication (Plahl 2020: 06:49). Robots taking over parts of the work that is usually done by caretakers naturally eases the burden on the latter and improves their performance (Deutscher Ethikrat 2020: p.8). Thereby, another positive side effect is being achieved, namely a surplus of time that caregivers can then reinvest into interpersonal relations with the care recipients (IPA320? 2011: 00:23) (Plahl 2020: 24:44).

Besides taking over physical tasks, certain systems, like robot companions, can serve as an interaction partner for care recipients, satisfying communicative and emotional needs (Deutscher Ethikrat 2020: p.19). An example for such a robot is the seal-like care robot Paro that reacts to recurrent sounds (Tansek 2020: 00:45).

Unlike the German Ethics Council, the documentaries put more emphasis on the entertainment function of robots. A reason for this could be that many robots that are currently in use in retirement homes are predominantly built for conversational purposes. This becomes evident when Pepper, a robot in the centre of many studies, lists its skills. Among other things, it is shown that it can play pairs, answer questions, sing and dance (WDR 2020: 01:06). These entertainment qualities allow for a more varied schedule in retirement homes (Regional 2017: 03:18) and are a welcome change for the care recipients (zeitpunktplus-Moma 2018: 03:49).

Another important aspect of the usage of care robots is the maintenance and, at least partial, restoration of autonomy, which is evident in both the statement of the German Ethics Council as well as in the documentaries. By using monitoring systems, one can achieve regular medical surveillance and social interaction even from a distance, as well as set reminders for the elderly to take their medication or drink sufficiently (Deutscher Ethikrat 2020: p.18). Humanoid assistance robots can enable care recipients to stay longer in their familiar home, allowing them to maintain social contacts that might otherwise be lost because of the relocation to a retirement home, and therefore reduce isolation (Plahl 2020: 24:44).

Restoring and maintaining the autonomy of care recipients also benefits their mental health, as mentioned by the German Ethics Council. For instance, technical compensation for lost abilities improves the feeling of self-efficacy as well as continuity (Deutscher Ethikrat 2020: p.35). Additionally, using auditory and visual stimuli to motivate dementia patients to react with touch, e.g. petting the care seal Paro, can elicit feelings of care and affiliation (Deutscher Ethikrat 2020: p.36).

3.1.2 Weaknesses

Even though there are a lot of strengths to be found when analysing the usage of care robots, both the documentaries and the statement of the German Ethics Council also state downsides. One of those weaknesses is that, socially, humans are still superior to robots (Tansek 2020: 02:43). The latter are limited in their cognitive and social abilities and unable to show empathy (BR Fernsehen 2019: 01:14) or to substitute human warmth (BR 24 2019: 01:18) and contact (3sat 2019: 04:17) (Deutscher Ethikrat 2020: p.51). They are also incapable of recognising and reacting to human emotions (Kroth 2018: 28:41) and do not possess emotions themselves.

In addition, they cannot fully cater to somebody's individual needs (zeitpunkt-plus-Moma 2018: 02:38), which is partly related to their lack of social skills, but also strongly linked to the acquisition, processing and storage of personal data (Deutscher Ethikrat 2020: p.47). The more personal data a robot system knows, the better it can help (Plahl 2020: 25:22).

Apart from being emotionally inferior to humans, there are certain limitations on their motor abilities that are predominantly addressed in the documentaries. For instance, in terms of motion and therapy, the state of the art is still not advanced enough to measure up to human skills (BR Fernsehen 2019: 03:03). Moreover, robots are incapable of distinguishing individual pills and medications, an important aspect of the everyday work in a retirement home (BR 24 2019: 01:28).

Another limitation accompanying robots, as mentioned by the Ethics Council, is their inability to take situational specifics into account. Caretakers still need to adapt the robot in specific situations (Deutscher Ethikrat 2020: p.42-43).

A weakness concerning the utilisation of care robots is the lack of funding for such a digitisation. Bearing in mind that large parts of the ongoing discourse address the role of care robots in counteracting the care crisis this aspect is of great importance. The German Ethics Council emphasises that high costs and many obstacles play into the development of such systems (Deutscher Ethikrat 2020: p.20). Even if these obstacles are overcome, there are further impediments to be managed, which becomes clear in the documentaries. For instance, robot assistance systems are not listed in the catalogue of health insurance funds (Kroth 2018: 17:01). Additionally, the digital upgrading of retirement homes with its associated costs, e.g. for additional staff like technicians and IT experts, barely receives any grants (Kroth 2018: 06:37). In the light of the underpayment of large parts of nursing staff the allocation of funds should be closely considered.

There is still a lack of understanding of AI systems in society. Care recipients might therefore adopt a negative position with respect to such systems, as they are unable to see the advantages they might bring (Main-Post 2019: 2:46) (WDR 2020). Furthermore, there can be unpredictable and unwanted consequences of technology (Deutscher Ethikrat 2020: p.31). For instance, care robots can help the elderly to maintain autonomy, an advantage mentioned earlier on. However, self-learning machines rather incapacitate people in need of care than serve them (Plahl 2020: 05:28).

As already mentioned earlier on, a great advantage of care robots is their ability of taking over multiple tasks and therewith relieving caregivers. At the same time, this can be seen as a problematic, as many tasks, e.g. body hygiene and lifting, benefit from an emotional connection which is not given by robots (Deutscher Ethikrat 2020: p.17). Many caretakers express their concerns regarding the possibility of robots taking on such intimate tasks. Due to their lack of social skills, robots cannot effectively communicate with the care recipients and might miss out on important information, e.g. concerning their well-being (Kroth 2018: 04:55).

Another problem is the data handling and storage. Individual care data needs to be saved and connected to the individuals. This data has to be accessible to several parties (Deutscher Ethikrat 2020: p.47). For example, when it is stored in a cloud, one has to trust the manufacturer of the service that it is not being misused. Thus, one has to place data privacy in the hands of an outsider. Overall, acquiring, processing and saving large amounts of data requires consideration of several aspects regarding data protection laws and, sometimes, even additional staff like external data protection officials are needed (Kroth 2018: 09:54).

3.1.3 Opportunities

As there are still additional functions which are not yet implemented but planned for the future, several opportunities arise. Many of them are similar to the strengths that have been described earlier on. On the one hand, this is due to the planned advancement of systems that are already in action. On the other hand, the actual usage of care robots in elderly homes is still very sparse and mostly experimental. With this in mind, the opportunities reflect the wishes and plans for the future.

As mentioned above, robots can lend support to caregivers. Since there is a lot of work that is either physically demanding or time consuming, there are also several tasks which might be taken over by robots. This relief still plays an important role for future developments. Currently, caretakers often have only limited time for listening and affection (Tansek 2020: 00:04). By supporting the nurses through taking on domestic tasks and chores that could easily be automated, robots could free up time, allowing to shift the focus from basic nursing to more active social interactions (Kroth 2018: 17:18) (Deutscher Ethikrat 2020: p.7). This again reflects the hope that with care robots, one has found the means to close the gap between the rising need of nursing and the shortage of nurses.

The surplus of time is not the only way in which the usage of care robots could improve the quality of care. The German Ethics Council emphasises that robots can also improve the sentiment of the elderly and relieve their stress directly, thereby improving their quality of life and decreasing the feeling of loneliness. Especially robot companions are well equipped for this task, as they could help to reduce feelings of loneliness, increase communication habits and form new connections by acting as an interaction partner (Deutscher Ethikrat 2020: p.19). This is also reflected in the documentaries. For example, some home residents are rarely visited. For them, the robot is like any other housemate (zeitpunktplus-Moma 2018: 04:30) and thereby, it could act as a substitute for missing visitors and social interaction. Additionally, the time a robot can dedicate to the elderly is mostly unlimited, and, taking the seal-like care robot Paro as an example, they stay in contact with the care recipients for as long as they want them to. As opposed to this, a living creature, e.g. a dog, might simply turn away their attention (Plahl 2020: 19:26).

Another strength of AI robots also offers one of the biggest opportunities for the field. By preserving the care recipients' independence through the use of the previously discussed monitoring systems, one could enable them to stay longer in their trusted homes. Such systems could allow for care and support from the distance, and by carrying out everyday chores, minimise the dependence on others. Thereby, they might ensure a longer, at least mostly, independent way of living (Deutscher Ethikrat 2020: p.17-18). This is also congruent with future visions, in which technology should, above all, serve those affected. Proper and good technology should support care recipients in their autonomy (Plahl 2020: 09:32).

Another interesting opportunity predominantly discussed by the German Ethics Council is the usage of robots for rehabilitation. They could help to compensate and activate everyday functions and promote plasticity, i.e. help to reconstruct lost abilities (Deutscher Ethikrat 2020: p.25). One applied example for that are computer-generated photo galleries that can stimulate impaired short-term memory (Plahl 2020: 10:05). Developing and deploying more such rehabilitative systems could restore, maintain and increase the quality of life for the elderly. In addition, these systems could help in recognising the need for assistance at an early point in time (Deutscher Ethikrat 2020: p.36). For example, in a game of pairs, the computer might automatically decrease the level of difficulty when the patient is not in their best form. For caregivers, this could serve as cue, indicating the changing health condition of the care recipient (Plahl 2020: 10:29).

3.1.4 Threats

The idea of robots as caregivers is not only sparking enthusiasm but is accompanied by worries, fears and threats for many people. All of these occur at different levels of the care work and are addressed by both the German Ethics Council and the various agents in the documentaries. Some of the mentioned threats are emphasised more in one of the two sources, while some appear in both.

Both sources name many similar aspects. One of the main threats pointed out by the Ethics Council and several of the agents within the documentaries is the danger of losing real personal relations and contacts. One of the documentaries highlights one particular aspect of personal contacts, which is the importance of physical contact in elder care and the potential danger of losing them (3sat 2019: 2:55 min).

The issue of data security, as already mentioned in the weaknesses, can be counted as a threat as well. If one takes the use of data to an extreme, it can potentially cause further loss of privacy. This is particularly threatening as this data might also include health data. A threat which is also discussed in other chapters such as in the chapter concerning medical diagnostic subsection ’Threats’. One example given in the documentaries is whether a robot is still learning when meeting his patients, in which case there is great potential for the robot’s company to gain data about the people (Kroth 2018: 28:41-29:20). Similarly, the German Ethics council is worried about the usage of the data by third parties (Deutscher Ethikrat 2020: p.23).

A threat that is voiced regularly when talking about robots in general is also discussed in relation to care robots. Both in the statement of the Ethics Council and in a statement of a member of the Ethics Council within a documentary, it is pointed out that there should always be the possibility to stop a robot in the case of a loss of control, meaning that there should always be a button to turn off the robot in case of an emergency (Plahl 2020: 23:35).

One fear that is heavily discussed publicly not only for this particular application for robots, but for many, is the fear that people will lose their jobs due to the use of robots. This threat to healthcare and care workers was discussed in both sources. The documentaries showed two different points of views on this topic, in which there was both fear and people not recognising this fear as relevant. Nonetheless, some care givers view the fear as unjustified, due to the different approach and quality in interacting with humans (zeitpunktplus-Moma 2018: 2:30 min).

An argument mentioned in the documentaries was that, apart from the differences in the interpersonal relationships, profit often drives the decision-making and could lead the decision towards the cheaper options, which might include robots in order to reduce paid workers in the future (Kroth 2018: 18:50-19:22). Related to this the Ethics Council and the documentaries noted the amount of funding for the development of robots for healthcare, which could influence such developments (Deutscher Ethikrat 2020: p.8) (Kroth 2018: 13:13).

The German Ethics Council further mentions infantilisation as a threat, as the usage of care robots only simulates relational behaviour and emotional connection (Deutscher Ethikrat 2020: p.19). Further, it might cause the feeling of humiliation in the caretakers. This could worsen, as the usage of robots could constantly confront elders with their limitations in everyday life. The danger of seeing a robot as a person is further discussed by the Ethics council. For example, a robot may be seen as a husband or wife. A disappearance of the robot due to a needed repair or a replacement of the robot could then possibly lead to a depressive episode (Deutscher Ethikrat 2020: p.38).

The documentaries named scepticism as another threat or fear, because a lot of scepticism might arise when hearing about the use of robots in healthcare, without being educated about the actual usage in place. It is mentioned, that often some of the scepticism can be reduced, when getting to know and experiencing care robots and their functions (Main-Post 2019: 0:14 min).

3.1.5 Demands

In contrast to the way the documentaries have portrayed the vision for the future of robots in elderly care, which included mostly hopes and wishes, the German Ethics Council has incorporated demands in their recommendations. In their official role to inform the public, respond to arising questions and review chosen topics from an ethical perspective, the council predicts consequences for individuals. To ensure the well-being of all those affected by new applications such as robots, it ought to propose rules and guidelines to avoid harm and problems. The requested demands can be seen as prevention of issues that might arise in the use of robots in the care sector. In the light of this project it is a useful indication of worries, fears or scepticism the public eye holds, with regards to robots in elderly care and even AI in general.

The entirety of demands within the German Ethics Council's statement can be further divided into separate categories by the issues they address.

One issue the council addresses is the accessibility of applications. The chance to use robots in a variety of ways in elderly care comes with the price of flexibility (Deutscher Ethikrat 2020: p.41). Machines and gadgets must be easy to handle for a variety of users and adjust to special needs individually (Deutscher Ethikrat 2020: p.10). In addition, everybody in need for help should have access to appliances, without being limited by financial constraints (Deutscher Ethikrat 2020: p.44). That creates a dilemma for the government. The additional costs must be taken into consideration but should not cause a deficit in other aspects of funding, for human staff or the improvement of the nursing crisis (Deutscher Ethikrat 2020: p.47) (Deutscher Ethikrat 2020: p.50).

Not only is the issue of funding to be decided by the legislative. There are more legal questions the German Ethics Council demands to be answered (Deutscher Ethikrat 2020: p.49). For one, the security standards must be set high and a structure for liability in the case of damage has to be established (Deutscher Ethikrat 2020: p.47-48). The existing guidelines must be adjusted to the potential new areas of application. Regular check-ups and preventive measures in production, use and authorisation are vital to ensure the well-being of care recipients (Deutscher Ethikrat 2020: p.34). With the use of technical applications in personal care the acquisition of data is inevitable and an issue the Ethics Council takes very seriously. To protect personal information, data security must be upheld to the highest standards as care is a deeply private and sensitive matter (Deutscher Ethikrat 2020: p.47).

The biggest legal issue that comes with the use of robots in personal care is the hierarchies of responsibilities. In the committee's statement this is addressed multiple times. A transparent structure of duties and responsibilities on a levelwise differentiation is required to protect care recipients and caretakers alike (Deutscher Ethikrat 2020: p.13) (Deutscher Ethikrat 2020: p.29).

To make sure the robots are used correctly and to avoid harm, caretakers must be given sufficient education in the use and operation. The analysed documentaries share this demand and request to incorporate the use of new technologies into the training system (Deutscher Ethikrat 2020: p.43-44) (ZDF 2019). But education and the spread of information may not end there. It is important to explain all implications of using various technological appliances, including robots and AI applications, to the care recipients and their dependants, especially before choosing a care environment (Deutscher Ethikrat 2020: p.32).

It is of essence to point out that the final application in care facilities does not mark the beginning of responsibilities. Starting in development and research for new applications, the cooperation with those affected is very beneficial. As researchers mostly consist of men implementing applications and women using them, a bridge to understand each other's restrictions and requirements can only support the development (Deutscher Ethikrat 2020: p.41-42). Although it must be emphasised that the elderly shall not function as a figurative guinea pig for researches, their testing and participation in research is a great aid. It helps to incorporate and account for their social norms and values while implementing (Deutscher Ethikrat 2020: p.41). Beginning with the first thought, ethical questions must be remembered in the process of development as well as other social sciences as these technologies affect the procedure of a very private matter (Deutscher Ethikrat 2020: p.39) (Deutscher Ethikrat 2020: p.42).

The German Ethics Council further proposes the differentiation of operational areas. Besides individual consent to allow the use of robots in personal care, a guideline is demanded to differentiate areas where robots can take on bits of the workload (Deutscher Ethikrat 2020: p.34). According to the Ethics Council's statement, other areas of nursing shall stay clear of their use to guarantee the maintenance of contacts. The ultimate goal is to provide the best possible nursing for care recipients to ensure their well-being, autonomy, security and protection of intimate and shameful aspects of personal care (Deutscher Ethikrat 2020: p.29).

In discussing the requirements for robots in care, not only economic and organisational issues are addressed in the source material, but they also listen to those directly affected by the use of robots in elderly care. (Deutscher Ethikrat 2020: p.49). The caregivers who work closely with the new applications must be able to operate them easily, so the attempt for improvements does not put more workload on them. Although care facilities should take the possible use of robotic systems and their dynamic development into account while planning care facilities, robots are seen merely as complementary elements to human care rather than as a substituting element (Deutscher Ethikrat 2020: p.13).

When talking about changes in elderly care, it is evident that the interests of the care recipients need to be a priority. The German Ethics Council poses demands to protect the care recipient's rights and needs (Deutscher Ethikrat 2020: p.42-43). In elderly care, the individual use of care applications like robots or AI must be decided by the person affected (Deutscher Ethikrat 2020: p.30). Every effort of aid should be personalised to the individual's needs and abilities and regularly checked upon to agree with the premise of good care and the person's well-being. In no way must the devices be used to manipulate or fool the elderly (Deutscher Ethikrat 2020: p.38).

It is stressed in the statement, that robots cannot replace true social interactions and all actions must be taken to prevent the loss of human connections (Deutscher Ethikrat 2020: p.36).

Overall, all demands made by the German Ethics Council aim to ensure the improvement of care quality in the use of robotics and AI (Deutscher Ethikrat 2020: p.50). It is to note that in this deeply personal matter one does not look for the technologically possible but rather the desired applications in care (Deutscher Ethikrat 2020: p.40).

3.2 Differences

In general, the two sources discuss similar aspects. As mentioned above, many strengths, weaknesses, opportunities, threats and demands are addressed in both the documentaries and statement. When it comes to weaknesses, there is a small difference between the sources when they talk about the lack of funding. The German Ethics Council touches more upon the lack of funding concerning the development while the documentaries rather discuss that, at this point, the care facilities have to pay for the systems themselves which leads to high costs. Generally, it can be observed that the German Ethics Council deals with the development and what could be possible in the future instead of what is possible or even the reality today. This is different in the documentaries as they show how some robots are used currently and how they contribute to a variety of tasks while also discussing what could be better in the future. The documentaries voice that up until now, robots are mainly used for entertaining the elderly and to add a degree of variety to their everyday life rather than helping with tasks like lifting, washing, etc. The German Ethics Council, on the other hand, does not really discuss how the robots can contribute to the entertainment in detail but focuses more on other tasks like lifting, monitoring and enabling care from a distance. With respect to the threats that are discussed, it can be observed that the German Ethics Council elaborates on the topic in a more detailed way and discusses everything from multiple points of view. However, the documentaries go one step further than the German Ethics Council with respect to the theoretical and practical considerations. Often, they already tested the AI for a while and discuss the real-life experiences, positive and negative ones, of the elderly and their relatives as well as the caretakers and computer scientists. The German Ethics Council reflects the demands and emotions of the parties concerned at some points but the documentaries show these directly by interviewing the parties.

The main difference lies in the way the respective source communicates their aspects to the outside world. The German Ethics Council is very objective and tries not to be opinion-forming while the documentaries differ in this aspect. Some documentaries are objective as well while others are rather emotional and try to trigger the emotions of the viewer, e.g. by using the title 'Hilfe, Wäscht mich bald ein Roboter?' (meaning 'Help, will a robot wash me in the near future?') (Main-Post 2019). Documentaries try to entertain and to catch the viewers attention whereas the Ethics Council wants to participate in politics. It is also relevant to note that the statement of the German Ethics Council only consists of written text without pictures or similar elements. The documentaries, on the other hand, consist of videos, speech and further media that may affect the viewer in a different way.

4 Discussion

Coming from an academic point of view with a background in Artificial Intelligence, the selected sources gave reason to select certain aspects for discussion. Be it misusing the label of AI, difficulties in communication or merely a lack of public understanding of Artificial Intelligence, the issues must be addressed and analysed to make the public perception and understanding of AI clear and accurate.

By analysing the selected source material, it has become apparent that the discourse between the legislative power and research in the department of robotics in elderly care seems rather limited. While the majority of the researchers introduced in the documentaries have raised the demand for guidelines and laws to frame the possibilities of their work, the German Ethics Council returns the inquiry. The council requests providers and developers of robotic technologies to take partial responsibility for the applications (Deutscher Ethikrat 2020: p.30). A clear need for an open, ongoing discussion consists to enhance collaborative work for optimal results.

Not only the consultation between the legislative and academic research appears to require improvement, but even more though the consultation between research facilities. Many of the examined videos put an emphasis on the robot Pepper and the opportunity of using it as a companion and entertainer for the elderly. In doing so the different researchers presented their individual work on games and software to get the robot ready for use. Rather than sharing intermediate results and accomplishments, to give each other a foundation to build upon, it appears as though they did not profit from each other and tried to accomplish similar results individually. This kind of working is not only an issue with research in Artificial Intelligence but rather connects to the need for Open Science at large. In this example it is evident how secrecy and competition clearly hinders the fast development of aid systems that could assist in solving one of the big societal issues of our time.

The continuing theme of mislabelling technologies as Artificial Intelligence is very prominent among the public display of robots in elderly care. For clickbait or merely for the purpose of entertainment, many analysed documentaries simplified the term AI for applications that do not fall under that definition. Additional lurid headlines purposely causing caution and fear were used to make the video more appealing. This not only caused distress in our research but further fuels the arbitrary, interchangeable use of the AI label with robots or advanced technology. Accusations of these proceedings leading to the misinformation and misleading of the public can be made. A more precise and careful use of the AI-label does not only benefit the understanding of Artificial Intelligence in the public eye but presumably reduces the fear of modern applications and increases the acceptance. If people had a clearer understanding of what to expect from new applications of robots in elderly care, using AI to operate, the scepticism and aversion could be reduced.

In the analysis and during the working process one striking theme occurred that was addressed only minimally in the selected source material but was quite noticeable coming from a background in AI nevertheless. As many robotic applications that incorporate Artificial Intelligence are still in their early days, it seems as though many worries and fears are directed against features that are not even functioning yet. The mismatch between the state of the art of robotics and vocalised concerns towards futuristic functions is severe. Looking even beyond the German care facilities to Japan where robotic applications are already incorporated deeper into the structure of elderly care, the premise of entirely replacing human personnel is nowhere in the future. With catchy titles roughly translatable to 'Reality: Robots instead of caregivers in retirement homes' (zeitpunkt-plus-Moma 2018) or caregivers fearing for their jobs, the public picture of AI is painted mostly wrong once more. Are these issues handled with necessary caution or does the discussion about potential issues of the future only stir up fear?

Many of these do not only occur when discussing robots in elderly care but in almost all areas of application of Artificial Intelligence. Once more the appeal has to be made, to dismantle the almost mythical nature of AI in our daily life. Without more education, the spreading of information and accurate descriptions, the fear and aversion may cause helpful resources to be limited by human ignorance.

5 Conclusion

In summary, one can say that both the German Ethics Council and the documentaries currently agree that although the robots do not have the potential to stop the nursing crisis in Germany yet, they could still be a helping hand. Furthermore, both sides called for politicians and legislators in Germany to put laws into place. So far, many of the agents we looked at waited for answers by respective politicians. However, we have to note that some of the documentaries were one-sided or could only show a small picture of the applications. We mainly obtained this broad picture of arguments from within the many documentaries that we looked at. Often, the short ones were focused on showing only a small picture of the applications, which reflects the knowledge of the agents of those documentaries. Even though this seems obvious, some videos excluded huge parts of the argumentation. Therefore, the statement by the Ethics Council often goes much deeper into the argumentation, while the documentaries are often based on personal experiences and opinions.

References

3sat. 2019. *Pflegeroboter helfen.* https://www.3sat.de/wissen/nano/pflegeroboter-nano-100.html.

Birthe Sönnichsen, ARD-Hauptstadtstudio. 2020. *Mein Helfer, der Pflegeroboter.* https://www.tagesschau.de/inland/pflege-roboter-101.html.

BR 24. 2019. *Caritas testet Pflegeroboter „Pepper“.* https://www.youtube.com/watch?v=UMWkSQz3aOo.

BR Fernsehen. 2019. *Wie ein Roboter in der Altenpflege helfen kann.*

Bundesamt für Justiz. 2007. *Gesetz zur Einrichtung des Deutschen Ethikrats.* https://www.gesetze-im-internet.de/ethrg/BJNR138500007.html.

Bundesministerium für Gesundheit. 2021. *Zahlen und Fakten zur Pflegeversicherung.* https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Statistiken/Pflegeversicherung/Zahlen_und_Fakten/Zahlen_und_Fakten_der_SPV_Februar-2021_bf.pdf.

Deutscher Ethikrat. 2020. *Robotik für gute pflege.*

IBM Cloud Education. 2020. *What is strong AI.* https://www.ibm.com/cloud/learn/strong-ai.

IPA320? 2011. *Serviceroboter im Altenheim: Care-O-bot 3 und CASERO.* https://www.youtube.com/watch?v=nJj8wJg6jNM.

Kroth, Isabella. 2018. *Schöne neue Pflegewelt? Digitalisierung im Altenheim.* https://www.youtube.com/watch?v=ZiUxr6R41xM.

Main-Post. 2019. *Zukunft der Pflege: Hilfe, wäscht mich bald ein Roboter?* https://www.youtube.com/watch?v=4jvsHum4zqU.

Meister, Jan Christoph, Marco Petris, Christian Bruck, Malte Meister, Marie Flüh, Jan Horstmann, Janina Jacke, Mareike Schumacher & Evelyn Gius. 2019. *CATMA.* Version 6.0.0. DOI: 10.5281/zenodo.3523228. https://doi.org/10.5281/zenodo.3523228.

OECD. 2011. *Help wanted? Providing and paying for long-term care.* https://ec.europa.eu/health//sites/health/files/state/docs/oecd_helpwanted_en.pdf.

Plahl, Silvia. 2020. *Roboter im Altenheim.* https://www.swr.de/swr2/wissen/roboter-im-altenheim-swr2-wissen-2020-08-19-102.html.

Regional, Sat.1. 2017. *Hilfe für Demenzkranke: Pflegeroboter "emma" bringt Schwung in Kieler Senioren-WG.* https://www.youtube.com/watch?v=9kjOKkDFEe8&t=2s.

Tansek, Tina. 2020. *Künstliche Intelligenz in der Pflege mit Pflegerobbe Paro.* https://www.youtube.com/watch?v=xgFUtgzlwzY.

WDR. 2020. *Pflegeroboter.* https://www1.wdr.de/mediathek/video/sendungen/video-pflegeroboter-100.html.

ZDF. 2019. *Der Pflegeroboter (8).* https://www.zdf.de/nachrichten/zdf-mittagsmagazin/die-pflegeklasse-teil-acht-der-pflegeroboter-100.html.

zeitpunktplus-Moma. 2018. *Realität: Roboter statt Pflegerin oder Pfleger im Altenheim.* https://www.youtube.com/watch?v=tQf_JEPwzKI.

Chapter 20

Asian reactions to AI supremacy in the game of Go

Sarah Neuhoff, Ralf Krüger & Nikola Tsarigradski

Go is a ancient game with a very intense player-base. For many Asian players, Go is the main part of their identity. In March of 2016 an AI, made by DeepMind beat the worlds best Go player Lee Sedol. The Go community did not expect this. We analysed mostly Korean newspaper articles to find out how people react when they are surpassed at their most valued skill. While surprise, fear and shock were part of the reaction, hope and optimism prevailed. We interviewed a professional Go player to get an up-to-date view on the situation. Since many of the reactions were connected to predictions about what would happen, we checked if these predictions held up. The main fear of job loss and general uselessness did not come true. Instead there was a great influx of new players and new ways of playing the game were discovered. We draw parallels to the real world and project our findings to the global scale. AI will augment human capabilities. For better or worse, depends on the humans.

Keywords: AI | AlphaGo | Go | Baduk | Asia | DeepMind

1 Introduction

For a long time, humans have competed against their own creations - machines. It has been about a century that mechanical calculators overtook brains at numeric computing ability. Ever since we have lost many races - we lost the physical race to cars, and now we also seem to lose the race about who is driving them better to AI. The losses, that hurt the most, however, are the ones, which are fought in domains we held dear due to our innate abstract cognitive abilities. If asked what is the most astounding skill of a human mind, perhaps language comes to mind, or the ability to comprehend and create art, or to sense and understand the feelings of another human. Being outperformed at these tasks would surely be a heavy blow to humankind, even though it seems that despite all AI advances we are still clear of any real competition.

The games of Chess and Go (kor. baduk) are domains, at which machines initially had strong difficulties performing. Partially due to that, humanity considers itself quite talented at playing them. They have the ruthless property that unlike car driving, medical diagnostics or text generation, we can directly match the machine's ability against the human's by having them play one another. The exact dates that mark the beginning of AI's reign over Chess and Go are the 17th of February, 1997 and the 11th of March, 2016 respectively, with DeepBlue and AlphaGo beating the world's strongest players at their time, Garry Kasparov and Lee Sedol.

What we find most peculiar about these events is, that AI's superiority over the human mind at the given domain became indisputable. As it seems to be a viable prediction, or concern, that AI will at some point outperform human ability at any given domain, it becomes a relevant question, how humanity will cope with this. Therefore, we intend to analyze the sociological effects of the AlphaGo match on the go community, treating it as a microsphere of human society as a whole. Based on that, we draw conclusions on the effects of superhuman artificial general intelligence on society, as well as required safety measures, in order to advert most drawbacks and alleviate the justified fears.

2 Game rules

This will be a short introduction to the rules of the game. If you have played go before, you can skip to the next section. If you want to read up on the rules in greater detail, see e.g. the wikipedia page of the rules. (Soojang 2021)

Go is usually played on a grid board of size 19x19, although occasionally also on smaller boards like 13x13 or 9x9. Players take turns placing black and white stones on the grid's intersections, with black always playing first. The goal of the game is to amass more points than the opponent, which can be achieved by surrounding empty territory on the board as well as capturing enemy stones. Stones that share a vertex connect to form one group. Diagonal stones are not connected. The connections to empty intersections of a group are called its 'liberties', and a group is captured, if all of its liberties are removed (see figure 1). One is not permitted to play in a manner, such that an own group would lose its last liberty (see figure 2).

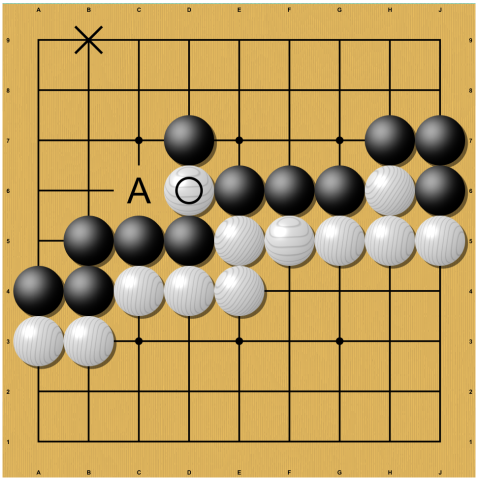


Figure 1: Atari - the marked white stone has only one liberty at A, a group with only one liberty is said to be in atari

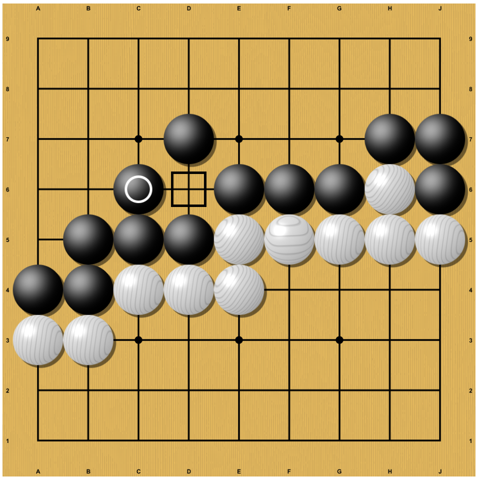


Figure 2: Capture - at the marked intersection a white stone got captured, white can't play on this intersection anymore, as they wouldn't have any liberties

When a player thinks, they don't have any good moves left to play they can pass. The game ends once both players pass. Afterwards the territory is counted and the winner proclaimed (see figure 3).

This is a rather simple example. Usually, within a game, countless complications and possible tactical motifs emerge, making open positions during a game

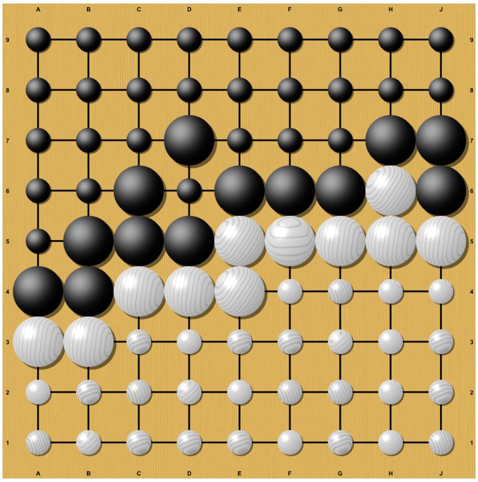


Figure 3: Counting - both players surround 29 points of territory, but black has captured an additional white stone, giving them 30 points. However, white gets 6,5 points of *komi* as compensation for black playing first, thus white wins by 5,5 points.

almost impossible to evaluate deterministically (see figure 4). Player strength is measured in ranks, kyu (k), amateur dan (d) and professional dan (p). Kyu ranks decrease, while dan ranks increase with strength.

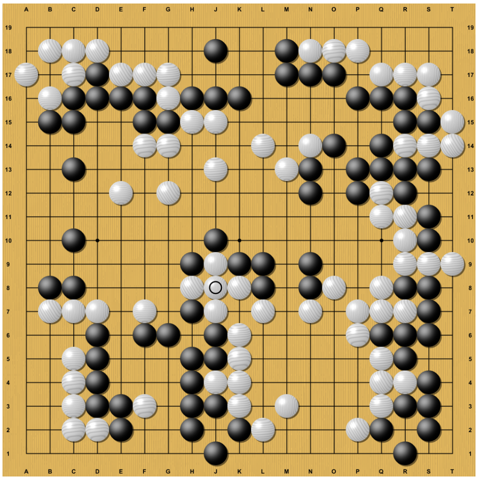


Figure 4: a game position from a real game between two dan players

3 Methods

In order to find appropriate sources we searched the internet for popular Chinese, Korean and Japanese newspapers, most of which also had an English version. We then filtered for keywords like ”alpha go”, ”go Ai” or ”baduk Ai”. In total, we found ten articles dealing with the match of AlphaGo against Lee Sedol or/and the consequences of the AI being better than any human player. Most of them were in the Korea Joong Ang Daily, but we also found a few articles in the South China Morning Post and one article from The Japan Times. Later we decided to also include some articles from non-Asian journals to have a bigger foundation for our analysis. We used the ontology that was developed as the basis of the book beforehand and picked those parts of it that were relevant to our topic 2. We decided to focus on attitudes and emotions, psychological effects, opportunity or threat and proposal of action. Then we read the articles closely and marked those sentences that dealt with one of the aforementioned categories. In order to collect the relevant information we created a table with HackMD with the columns ”title” and then the four categories. For each article we wrote quotes in the respective column and which tag of this category fitted. Some of the tags we used are not part of the ontology but are relevant in the articles, so we decided to include them. We added ’frustration’ for the emotion & attitudes category, ’hurt confidence of human players’ for the psychological effects category, ’inspiration for Go’ for the opportunity category and ’end of human culture or species’ and ’loss of emotion’ for the threat category.. Afterwards we counted in how many articles each tag occurred in order to see how relevant it was. We coded the occurrences in the following way: 7-9 articles: very often, 5-6 articles: frequently, 3-4 articles: sometimes, 2 articles: rarely, 1 article: once. When using tags explicitly during the following results and discussion they are surrounded by speech marks to make clear that the phrasing is adopted by the ontology.

4 Results and Discussion

4.1 Emotions and Attitudes

What we find when looking at the frequency of different emotions in the articles is that ”surprise” is by far the most common. To be more specific it is primarily a negative surprise namely shock. A likely reason for this is that at least in Asia the general public did not took AI seriously until AlphaGo was created and suddenly there was an AI that was better than any human player. Nobody had anticipated this result, not even Lee Sedol himself (Kohs 2017). A few months before there

Table 1: results

Emotions and Attitudes	#
Frustation	2
Hope	4
Fear	5
Doubt	3
Surprise	9
Uncertainty	2

Psychological Effects	#
May increase anxiety towards job insecurity	5
Hurt confidence	2

Opportunities and Threats	#
Inspiration for Go	6
Suitable way to solve unsolved problems	2
Optimization of complex systems	2
AI kills human jobs	5
End of human culture or species	3
Loss of emotion	2

Proposals of Action	#
Increase investment/research in AI	5
Social and economical restructuring and reformation	6
Information/education about AI	3
Ethical guidelines	2

was already a match played by AlphaGo against a European Go champion, which AlphaGo won. On basis of seeing the algorithm's performance during this match Sedol concluded that the algorithm was beatable. One of the reasons it was such a huge surprise is that the AlphaGo team greatly improved the algorithm in between the European and the Lee Sedol match. Before the first match, Sedol was so sure to win, he said: "I don't think that it will be a very close match. [...] My hope is that it will be either five-zero for me, or maybe four to one." (Kohs 2017). The general opinion was that he would easily beat AlphaGo. But then it turned out entirely different, the match ended with a 4 to 1 win in favour of the AIcite (Jong-soo 2017). The public was very shocked when it became clear how strong AI already was. This elicited not only shock but also even if rarely mentioned frustration in the professional players, because they were by no means prepared for losing (Peng 2017). According to one article Lee Sedol himself described his defeat as a "horrible experience" (Japan Times 2017).

Another emotion that occurred frequently is "fear", however what exactly people are afraid of is not often mentioned. One author is afraid of "the colder world"(Ki-chan 2016) that our future might become. Seemingly he worries that the greater use of AI will lead to a lack of emotion and empathy. Another author thinks that an AI revolution will give power to only a few (Zastrow 2016), probably already rich and powerful people, who have control over the AI, while the rest of humanity is subordinate to them. The gap between those who have power and those who do not would get even wider.

On the other hand "hope" is mentioned almost equally often. The authors who mention hope believe that Go AI will give human Go players inspiration for their way of playing, that AI can show them new ways how to play Go and foster their creativity. Also Go AI provides new interpretations of the game (Peng 2017). One author also writes about the general hope that AI will at some point be better at certain tasks like advising or teaching than humans and therefore can provide qualitatively better services (Lee 2017).

Another emotion, though its occurrence is rare, is "uncertainty". This uncertainty is also described as "uncomfortable" (Kee-eung 2016) and a consequence of the fact that one can not know what Alpha Go is thinking, because it is not really thinking at all. Nevertheless it unsettles people that there is a smart agent, but they can not trace its line of thought. As one author put it: "Even for the masters, the inner center [of AlphaGo] is a realm of the clouds, murky and unfathomable" (Seung-il 2017).

An attitude sometimes occurring in the articles is "doubt". More specifically on the one hand it is general concern about the safety of AI. The authors are aware that there is still a lot to do in order to make AI safe enough to use it on a larger scale. Additionally, there is doubt about the skills of AI. Some authors believe that AI is unable to understand beauty and can not appreciate a beautiful move (Mollard & Roux 2016). Also they see limitations of AI in that it can not explain its own thinking and cannot cooperate or comprehend abstract concepts, which are all skills that humans have though (Japan Times 2017).

4.2 Psychological Effects

After Lee Sedol's defeat, it seems that professional Go-players have developed a fear of losing their jobs, as the tag ”increased anxiety towards job insecurity” occurs frequently in the articles. This is understandable, as an AI playing Go better than they do makes them feel dispensable. However fear of job loss is also mentioned with regard to the general population. So even people that do not have anything to do with Go during their work fear that AI will soon be so powerful that it will render them obsolete.

Another psychological effect that occurs rarely is the ”hurt confidence” of human players in their proficiency. Only one author commented on Lee Sedol's defeat that it ”felt as if human talent was being snatched away by computers” (Ki-chan 2016). This was a bit surprising, because we had thought that this feeling of needlessness, of being nothing against an AI, would occur more often. It feels like an intuitive reaction to being beaten. Furthermore the match gained really huge attention in Korea and the narrative at the time was that Sedol represented humankind. It became a very symbolic display of man vs machine, which weighted the defeat of Sedol even stronger (Kohs 2017).

4.3 Opportunities and Threats

When looking at the opportunities and threats people see regarding AI in general, it is striking that at least in the articles we analysed the outlook to the future is very balanced between positive and negative, with a slight lean towards the positive. ”Inspiration for Go” and more creative Go games are frequently mentioned opportunities. One author for example states that ”Go players can acquire new skills and make their contests more interesting” (Japan Times 2017) with the emergence of Go AI. The general opportunities of AI are mentioned less often, but this makes sense as the articles primarily deal with Go and AI. Rarely the authors wrote that AI will be able to solve problems for us that we can not solve ourselves. Also it is rarely mentioned that AI can optimize complex systems for us like ”transport passengers more efficiently, reduce traffic jams, avoid accidents, eliminate road rage” (Lee 2017).

Among the threats, ”killing jobs” is by far the most prominent and occurs frequently. A few authors are more dramatic and fear the ”end of human culture” or even the end of the human race, because AI would soon render humans completely useless. One for example quotes Elon Musk: ”Humanity’s position on this planet depends on its intelligence. So if our intelligence is exceeded, it’s unlikely that we will remain in charge” (Lee 2017). Another threat that occurs rarely is the already before mentioned ”loss of emotions”.

4.4 Proposals of action

So what consequences should Korea and other Asian countries draw from this? One proposal that occurs frequently is an ”increase in research and investment in AI”. It seems that only after the historic match with Alpha Go, Korea realised how powerful AI is and that it will be an enormous chance for economy (Park 2016). Some authors even claim that the Alpha Go match started the fourth industrial revolution in Korea (Ki-chan 2016). Here, the significance of the event becomes clear, it really opened Korea’s eyes about AI and since then Korea has become a big player in AI. The other important proposal that is made frequently is ”social and economical restructuring and reformation”. It is proposed that theories on productivity and economics have to be changed such that the human workers can remain competitive against AI (Ki-chan 2016). Furthermore it is important to establish re-education for those workers who lose their job (Young-seok 2016). Job loss will concern many workers and one can not simply assign them a new task,but instead they need the time and resources to learn something new. Also a general education reform is demanded, although the author does not specify what exactly he means by this (Young-seok 2016). Another author stresses the importance of changing our way of thinking about AI. Concerning the job market we should not think in terms of ’AI against humans’, but more in terms of ’AI works for humans’ (Li 2020).

Sometimes it is also mentioned that education and information about AI should be fostered. However, the need for ”ethical guidelines” for AI is only rarely mentioned and is not the focus of the articles, probably because playing Go does not seem to be very closely related to ethical questions. One could imagine that this would be different if we were talking about an AI being better at diagnosing cancer for example.

5 Professional's insight

We felt it was a bit out of touch with the actual community, whose reaction and feelings we wanted to grasp, to only refer to newspaper sources. Hence we reached out to Finnish professional go player of the Japanese go association Antti Törmänen, 1p. We hoped to gain an insight into the perspective of someone actually immersed in the community we want to observe.

He confirmed the notion of shock and surprise that accompanied most articles. He says he expected a clear victory from Lee Sedol, as did most experts. His observations from within the professional community confirm the notion of job anxiety. That affects on the one hand the job of professional players themselves, as well as the teaching jobs many of them have. Antti himself says, he expected his student numbers to drop, given that publicly available superhuman AI could be used to replace a human teacher, however this fear didn't seem to come true. He stresses that the emergence of strong AI is a great opportunity for the go community. He himself uses AI for his studies, as without giving insight, it can quickly verify or refute ideas on the board. Additionally, having such strong teachers available makes the game itself more accessible to the public. More as a drawback than a threat, however, he emphasizes the currently largest problem with strong AI from his point of view - AI-assisted cheating. He even goes as far as to say that this drawback levels, if not outweighs the upsides.

6 Conclusion

We wanted to observe the Go community to see how they reacted when in 2016 their main identifying skill was mastered by AI. We have seen that there are many different reactions, ranging from the fear of culture being destroyed, to the hope that humans will be able to learn ever more with the help of AI.

What was most significant, however, was shock. No one in the Go community expected this. Go seemed to be beyond logical thinking, beyond what an algorithm could learn. From an outside perspective, this seems like a bold thing to claim about anything, but it is a claim that humans repeatedly make about a lot of *intuitive* human abilities. We claimed that chess was out of reach for computers, we claimed that Jeopardy was out of bound for computers and more recently we claimed that text generation was too difficult. In each case, we were proven wrong (Deep Blue, Watson, GPT-3).

The event certainly has awakened the Go community with a big scare. As chess players before them, they feared for their jobs. But surprisingly, most also saw the benefits. They hoped the strength of AI would help them learn more and the scale of the event would bring growth into the community.

We see in the Go community a great example of how people react to being beaten. with the benefit of hindsight, we can now evaluate whether the reactions were justified.

Strikingly, the thing most feared, job loss, did not occur. While Go was declining in popularity in Korea before, through this event, more people are now interested again (Jong-soo 2017). Teachers now have to compete with AI teachers, but the increase in students has also filled offline-classrooms. It is unclear how long this trend will continue.

The negative consequence of players using AI to cheat, which was brought to our attention by Antti Törmänen, was not mentioned in any article. This shows, that there might always be side effects of AI that we will not anticipate.

Beside the influx of new players into the community, the other hope that AI will help humans become better has also become largely true. To have a tool to evaluate game states instantly has been described as very beneficial for personal improvement. And to play against the AI, has shown many players new ways of thinking about the game. One player, Ke Jie, after playing (and losing) three games to AlphaGo, has analysed his losses and won the next 12 professional games against humans (Li 2020).

Now, what can we learn from this? What can we say about other domains in which AI will dominate us in the future? Mainly that predictions about the future based on emotions should not be trusted easily. Fear is not a good predictor of what the future holds. We saw that the real drawbacks of AI are when humans use it to cheat in systems of fair competition. in the real world, there is no cheating, but global unfairness still might be on the rise with AI supported unfair humans. But we have also seen hope come true. with the help of AI more people will be able to get the benefit of learning, of getting excited about interesting topics. More and creative ways of living might be shown to us by AI. In the end, AI is a tool to magnify human capabilities. for better or worse, depends on the humans.

References

Ki-chan, Kim. 2016. Ai versus humanity. *Korea JoongAng Daily*. https://koreajoongangdaily.joins.com/2016/03/28/columns/AI-versus-humanity/3016769.html (1 March, 2021).

Japan Times. 2017. AlphaGo AI stuns go community. *The Japan Times*. https://www.japantimes.co.jp/opinion/2017/06/03/editorials/alphago-ai-stuns-go-community/ (1 March, 2021).

Jong-soo, Sohn. 2017. Go in the 21st century. *Korea JoongAng Daily*. https://koreajoongangdaily.joins.com/2017/03/22/columns/Go-in-the-21st-century/3031303.html (1 March, 2021).

Kee-eung, Kim. 2016. Lessons from the AlphaGo shock. *Korea JoongAng Daily*. https://koreajoongangdaily.joins.com/2016/03/15/columns/Lessons-from-the-AlphaGo-shock/3016258.html (1 March, 2021).

Kohs, Greg. 2017. *Alphago - The movie | full documentary*. Moxie Pictures & Reel as Dirt.

Lee, Newton. 2017. Alphago's china showdown: why it's time to embrace artificial intelligence. *South China Morning Post*. https://www.scmp.com/week-asia/society/article/2094870/alphagos-china-showdown-why-its-time-embrace-artificial (1 March, 2021).

Li, Michael. 2020. No human being can beat Google‘s AlphaGo, and it's a good thing. *hackernoon*. https://hackernoon.com/no-human-being-can-beat-googles-alphago-and-its-a-good-thing-9u7r36b8 (1 March, 2021).

Mollard, Pascale & Mariëtte Le Roux. 2016. Game over? New AI challenge to human smarts (update). *Phys.org*. https://phys.org/news/2016-03-game-AI-human-smarts.html (1 March, 2021).

Park, Yvonne. 2016. The AlphaGo challenge. *Korea JoongAng Daily*. https://koreajoongangdaily.joins.com/2016/03/17/columns/The-AlphaGo-challenge/3016370.html (1 March, 2021).

Peng, Tony. 2017. AlphaGo: The ultimate Go master. *Synced*. https://medium.com/syncedreview/alphago-the-ultimate-go-master-7008ac4ce488 (1 March, 2021).

Seung-il, Hong. 2017. Learning from AlphaGo. *Korea JoongAng Daily*. https://koreajoongangdaily.joins.com/2017/01/08/columns/Learning-from-AlphaGo/3028376.html (1 March, 2021).

Soojang, Kim. 2021. *Korean 2016 Rules*. http://home.snafu.de/jasiek/k2016.html (1 March, 2021).

Young-seok, Park. 2016. Exploiting the AlphaGo shock. *Korea JoongAng Daily*. https://koreajoongangdaily.joins.com/2016/03/24/columns/Exploiting-the-AlphaGo-shock/3016630.html (1 March, 2021).

Zastrow, Mark. 2016. How victory for Google's Go AI is stoking fear in South Korea. *New Scientist*. https://www.newscientist.com/article/2080927-how-victory-for-googles-go-ai-is-stoking-fear-in-south-korea/ (1 March, 2021).

Part IV

Politics, governments and non-government organizations

Chapter 21

AI made in Germany

Christian Burmester, Thiago Goldschmidt, Felix Naujoks & Tom Pieper

Artificial intelligence (AI) has become one of the defining technologies of our time. Its increasing influence on our daily lives and the shaping of society makes it paramount for governments to have an appropriate strategy in place to address the risks and opportunities that it entails. We analyzed the AI Strategy of the Federal German government and tried to find answers to the following questions:

Which areas are mainly targeted by the AI strategy of the German Federal government? Does the strategy address the important risks and opportunities in an adequate way?

Our results suggest that the government is generally aware of the possibilities and risks of AI and proposes a plethora of measures to address them appropriately. Due to some shortfalls, however, it remains questionable if the proposals will prove sufficient enough to be meaningful long-term solutions.

Keywords: AI-made-in-Germany | German AI Strategy | society | research | economy

1 Introduction

Upon returning from a visit to China, where Angela Merkel was confronted with China's prodigious and well-advanced AI program, the German chancellor instructed her own government to commence a thorough assessment of Germany's standing concerning AI (Merkur.de 2018, Produktion.de 2018). This analysis resulted in the government publishing their own strategy in 2018: "AI made in Germany: Strategie Künstliche Intelligenz der Bundesregierung" ("Artificial Intelligence Strategy").

"Made in Germany" stands for reliable engineering since 1891 (Eduard Langwieser 2021), which has experienced great trust around the globe. With their strategy, the German government expands the label to fully harness the possibilities of AI.

The first purpose of the strategy was to consult with experts and decide on an agenda on how to embrace this technology in the years to come. The measures of the government aim at keeping Germany relevant as an economic power and therefore counteract the power concentration of current technological advancements.

Secondly, the strategy was written and published to inform the German public, as well as the international audience and possible partners about the national intentions regarding Artificial Intelligence. Besides the broad public, the strategy especially addresses researchers, scientists and companies to advise them about the financial and logistical support the government planned and plans on giving. Concurrently, the government discusses ethical and sociological impacts of AI on the public with the intention to foster a nationwide discourse leading to a more informed and reflective view on AI.

Initially, the government allocated 3 billion euros to implement the strategy. The budget was later increased to 5 billion euros in June 2020 (Deutsche Bundesregierung 2020). Who will be benefiting from this and if there are tendencies towards a certain area of society, how could this affect the public discourse on AI?

2 Definition of AI

2.1 Definition

In the preface of the paper, the German government defines the type of Artificial Intelligence. After stating that there is no single true definition, the government elaborates on its understanding of AI. The focus lies on weak AI, i.e. different methods of mathematical optimization for confined processes (Federal German Government 2018: 4). The mentioned fields and methods aligning with the ontology, created in this seminar, are automated reasoning, knowledge representation, machine learning (pattern recognition, human-machine-interface and understanding of human behavior) and the field of robotics. The government's definition demonstrates a well-informed and differentiated view of Artificial Intelligence which aligns on many levels with our ontology.

2.2 Digitalization vs. Artificial Intelligence

As discussed in the previous chapter, there is no single true definition of it, even though the German government specifies what AI is. The German government mixes opportunities/applications of digitalization in general, with the offered opportunities/applications of AI (Federal German Government 2018: 4,17). The definition of AI combined with the lack of a clear distinction of digitalization and AI by the German government could confuse readers.

3 Methodology

3.1 Quantitative approach

We decided to address the following questions utilizing a quantitative approach:

1. To which extent are the three key areas in our society targeted by the strategy?
2. What language is used when addressing those different areas?

To answer the first question, we agreed on three key areas. We allocated each relevant statement, measure and goal to its corresponding area using annotations. We extracted the annotations for each area and divided the total amount of words in the main document by the number of words the extracted file contained. This gave us a percentage that indicated how much of the text targeted which area.

Since the strategy also had 38 clearly defined goals, we decided to allocate them too, in order to add an extra layer of information. By dividing the amount of goals per area by the total amount of goals mentioned, we were able to tell their distribution in percentage.

As the tone of voice is a rather interpretable measure, we examined the key areas for prominent key words or phrases in order to find significant differences in the use of language. To get a comparable measure we ran a python TF-IDF (term frequency - inverse document frequency) package called SciKit Learn feature extraction. TF-IDF computes the relative relevance of words for a document in relation to the overall appearance in a set of documents. We treated the previously marked areas as separate documents to obtain the desired comparable frequency. With such a small number of documents, TF-IDF can be rather imprecise. Therefore, while stop words like pronouns, conjunctions etc. were filtered automatically by the algorithm, we still performed some manual post-processing sorting.

After sorting out all words contained in the 20 most frequent words of all three subcategories, we examined the rankings for patterns. While TF-IDF may not be the optimal solution to obtain a meaningful answer to what is focused on in the selected areas, for it being a more precise count of frequency, it can still give an insight into which possible keywords the German government emphasizes on.

3.2 Qualitative approach

In order to examine if the government is aware of all the positive and negative implications of AI, we looked into the strengths, weaknesses, opportunities, threats (SWOT) and fields of application identified in the ontology and compared these to their corresponding parts in the strategy.

As a first step, we compared the results of the SWOT analysis with the government's statements. We analyzed the actions and measures proposed by the government and assessed whether or not they matched those in the ontology.

To interpret our results and deduce implications, we agreed upon the following assertions.

If a strength of AI or an opportunity was not picked up on by the government, we interpreted it as a potential ignorance or lack of knowledge. This could have a distorting effect on the public's perception of AI. On the contrary, if all the strengths and opportunities were presented, we deemed it to be adequate.

If a weakness of AI or a threat was left out, we again deduced negative implications for the perception of AI. If, for example, a threat was not mentioned in the strategy, one could argue that the government is either not aware of it or does not show a sensible and reasonable response to it. Both cases would entail negative effects on the public's perception.

We looked at the fields of applications and the proposed actions in a similar manner. If any actions were not mentioned, we tried to discern what this could mean for the public discourse.

4 Quantitative approach

4.1 Amount of text and goals dedicated to each area

We posed the question "To which extent are the three key areas in our society targeted by the strategy?" to find out if the government's strategy fairly addresses and represents the whole population or if there were clear tendencies towards one subsection of society over another.

We divided society into three distinct groups of interest: economy, research and the public. Each of them requires its own set of goals and measures to address the opportunities and threats specific to their domain. It is crucial for a strategy to strike a balance when addressing those areas in order to fairly target the whole society.

Economy includes everything related to supporting businesses ranging from start-ups to big corporations and promoting actions that would enable Germany to become a relevant player in the international AI market.

The area of research emcompasses everything related to academic studies and the promotion of scientific advancements related to AI.

The public entails all measures and goals that are supposed to benefit the whole society. This ranges from regulations that safeguard privacy and security to ensuring that AI is developed in a human-centered way.

When analyzing the word count, we found out that approximately 55% of the document addresses the public, 46% was aimed towards research and 34% was dedicated to the economy. The sum of those numbers does not equate to 100% because a large portion of the text can be attributed to multiple areas simultaneously.

We are fully aware of the care we have to take when drawing conclusions solely based on word counts, since it completely neglects the meaning of the content.

To mitigate this issue, we decided to add an extra layer of information by analyzing the number of goals the government defined, dedicated to each area.

Our results support the notion that the public is the government's main focus since 58% of the goals are dedicated to it. In contrast to the results of our word analysis, a lot more goals are dedicated to economy (42%) than to research (24%).

This might be indicating that the government's priority is to make sure the public is prepared for the changes that lie ahead. The discrepancy between the results we achieved in word count analysis compared to the goal count analysis regarding research and economy, illustrates the only relative reliability of such a simple quantitative analysis for complex papers.

4.2 TF-IDF

To get a more differentiated and comparable view on the tone of voice, that the government communicates the measures and goals of the strategy with, we pursued the frequency measure TF-IDF. The algorithm constructs a ranking of the most frequent words per document (in our case per area). This allowed us to obtain a more detailed view on not only how much of the text is dedicated to an area, but also which specific words are emphasized in each area.

The sections concerned with economics and research share 4 of the most used words which are completely absent in the field of the public („Deutschland“: Germany (engl.), „Europa“: europe, „Wissenschaft“: science, „Forschung“: research). This emphasizes the importance the government places on the transfer of the knowledge acquired by AI research into concrete economically viable applications. The frequent use of the word „Europa“ suggests a willingness of international/European cooperations in research and economics.

With a very high frequency in the area of economics, the word “Mittelstand” poses an outlier. This stresses the government’s acknowledgement of KMUs (“Kleines mittelständisches Unternehmen”: small companies residing in the Mittelstand (engl.)) and their indispensable role for the German economy. It furthermore can be interpreted as a signal of support to smaller companies and KMU’s by the government.

To support this, 3 of the Top 20 most frequently used words in the field of the public are concerned with the term “application”. This suggests the endeavor of the German government to acquire the approval of the public by focusing on the implementation of AI into the daily lives of society.

4.3 Comment

To further address to which extent the three key areas are targeted, an analysis of the actual spending could yield a deeper insight. It is valuable to evaluate the planned measures, but a comparison of the money spent in each area could indicate where the German government currently sees the biggest impact induced by their monetary means.

5 Qualitative approach

5.1 Strengths

To be able to shape and execute the expansion of a technology, a government needs to be aware of its key strengths and weaknesses. Misinformation or lack of understanding could lead to over- or underestimating the opportunities and threats the technology offers. In this chapter, we analyzed the level of awareness of the German government concerning the strengths and weaknesses of AI and its plans to utilize and mitigate them.

5.1.1 Scalability

Scalability expresses AI's ability to be copied or multiplied with low additional costs. Computational resources can easily be added to scale up processing power and capabilities.

By suggesting measures like collaborative used testbeds (Federal German Government 2018: 15) or pre-designed AI systems developed in "Mittelstand 4.0 centres" that can be used universally by KMUs (Federal German Government 2018: 22) the German government demonstrates an implicit awareness of AI's scalability. Nonetheless, an explicit mentioning is missing.

5.1.2 Complexity

Complexity describes that AI is able to discover patterns that are too complex for humans to find or interpret. This is mainly due to its capability of handling enormous amounts of data in a short period of time.

The German government sees great potential of using AI to solve hard problems and enable breakthroughs in highly complex research areas, such as mobility, energy systems, agriculture, food security, healthcare, the protection of resources and mitigating climate change. The acknowledgement of possible research fields is a strong indication that the German government is aware of this strength of AI (Federal German Government 2018: 17).

5.1.3 Ability

Ability describes the potential of AI to automate human manual and cognitive tasks. Especially in repetitive tasks, AI often outperforms humans in terms of speed.

The German government clearly demonstrates its awareness of AI's ability to automate human intellectual tasks, by stating "[...] delegating monotonous or dangerous tasks to machines so that human beings can focus on using their creativity to resolve problems" (Federal German Government 2018: 25). The German government mentions recruiting in HR to exemplify the kind of task such systems could help us with. To bring up AI's potential to assist humans in such intellectual tasks, through managing applications and pre-selecting candidates, is a clear demonstration of the government's awareness of AI's ability (Federal German Government 2018: 28).

5.1.4 Consistency

Consistency describes that AI-Systems avoid typical human mistakes and provide reproducible results.

The German government elaborates on the increased safety, by reducing human mistakes, in today's mobility sector. By introducing autonomous vehicles and pointing out their vastly superior ability of reacting more precisely and reliably than a human agent, the government demonstrates its understanding of AI's consistency (Federal German Government 2018: 35).

5.2 Weaknesses

5.2.1 Explainability

Explainability describes AI's negative tendency to produce unpredictable or unexplainable results due to the inherent complexity of such systems. This lack of transparency and understanding leads to AI being perceived as a "black box".

The government seems to be completely aware of the importance of widespread understanding of AI and its underlying processes. It mentions that public resistance to AI resulting from a lack of explainability could prove a massive obstacle for innovation (Federal German Government 2018: 16). The government responds to this issue with a variety of measures, aiming to find and disclose the underlying processes to the broad public.

5.2.2 Computationality

Computationality depicts the property of AI to produce outcomes exclusively based on mathematical computation and reasoning. Thereby, AI disregards any context, human emotions, and awareness of situations leading to a strong vulnerability to biases.

While aiming to implement AI in more areas and transferring more responsibility to it, the German government is aware of this weakness and its possible negative impacts. These comprise not only of possible discrimination originating from biases, but also of super-rational and therefore possibly suboptimal decisions made by AI (Federal German Government 2018: 39).

5.2.3 Limitations

Limitation of Artificial Intelligence includes the enormous amount of training data and its quality needed to train an AI system. Additionally, the actual limits of AI depend strongly on its underlying model.

The limitation proposed by the amount of training data needed and its quality is a frequently addressed issue in this paper. The government proposes a lot of actions to overcome this restriction in the advancement of AI (Federal German Government 2018: 6,36).The limitation of computational power and the underlying models are not addressed specifically.

5.2.4 High Cost

The High Costs Artificial Intelligence entails are the enormous computational power needed by an AI system resulting in high energy consumption. Furthermore, a high expert knowledge and extensive access to data sources is needed to implement and work with AI.

The examination of the high costs of AI systems is mainly cut down to the high expert knowledge and the data sources needed. Those factors are addressed very directly with a lot of measures in the paper. The high energy consumption (Federal German Government 2018: 20) and the needed computational power (Federal German Government 2018: 32) are mentioned but not discussed any further in the strategy (Federal German Government 2018: 30).

6 Opportunities

AI will be one of the defining technologies of the next decades, opening up a long list of opportunities. Advancements in AI could enable us to tackle urgent challenges that humanity is currently facing and help us get a better understanding of our own intelligence. This is only possible when AI is leveraged to its full potential. In the following, we have analyzed if the 4 opportunities, presented in the ontology, are acknowledged by the government. We also looked into possible measures the government proposes in order to benefit from AI.

6.1 Productivity increase

When compared to humans, AI systems are able to fulfill tasks more efficiently and cheaply. If certain tasks are transferred to AI systems, it could free up valuable time for people to do more creative things.

6.1.1 Acknowledgement and measures

The government acknowledges the huge potential of AI to increase productivity in all areas of society (Federal German Government 2018: 25). It also picks up on the effect that this may have on the daily life of individuals. Repetitive and dangerous tasks may be operated by AI systems in the future, enabling society to focus on creative problem solving (Federal German Government 2018: 25). It also addresses the fact that, when using AI, certain processes or operations could be run much cheaper in terms of costs and energy consumption (Federal German Government 2018: 19).

To gain from these potential benefits of AI, the government proposes a number of measures. Firstly, it wants to promote experimental spaces for employees to establish direct contact with AI (Federal German Government 2018: 26). Secondly, an AI observatory shall monitor the expansion of AI and its socio-economic effects on the public (Federal German Government 2018: 26). Finally, the government wants to strengthen the expertise in research areas and facilitate a regulatory process that allows AI systems to get to the implementation phase sooner (Federal German Government 2018: 19).

6.1.2 Conclusion and criticism

The government is aware of this opportunity and presents ideas as to how an optimal productivity increase could be obtained. A good example of this is the government's program to educate and train people about AI who are likely to be dealing with AI in the future. For the implications of AI on our workplaces, however, it fails to conceive how to actively profit from the advantages that AI may bring. The plan to put into place an observatory to check and review trends and consequences of AI seems rather reactive and reserved at second glance. The strategy also lacks a long term view on how a society without many of today's jobs would look like and what could be necessary measures in order to move towards it (e.g. universal basic income (UBI)).

6.2 New kinds of problem solving

The advancements of AI provide humanity with new means to approach unsolved problems as well as optimising already solved but still very complex ones. AI is able to achieve this because of its ability to extract meaning from very large amounts of data.

6.2.1 Acknowledgement and measures

The first aspect mentioned by the government is the possibility to speed up medical procedures by deploying AI for earlier recognition of diseases and risks. In the same token, the government states that AI can even find new approaches or ways to detect diseases in the first place (Federal German Government 2018: 36). The government plans to build a network that shall cultivate the exchange and communication between scientific research and the healthcare sector (Federal German Government 2018: 37). In the mobility sector, the government acknowledges the potential of AI to make processes more efficient (Federal German Government 2018: 35). The plan is to build an infrastructure to facilitate the inter-connectivity of road users and the optimal flow of mobility data (Federal German Government 2018: 35). Another opportunity, that the government shows its awareness of, is the utilisation of AI for IT-security where large amounts of data need to be scanned in split seconds to detect anomalies (Federal German Government 2018: 18). The government wants to arrange for public research in the affected areas (Federal German Government 2018: 18).

6.2.2 Conclusion and criticism

This opportunity is well covered and appreciated by the government. It addresses the importance of using big data to solve demanding and complex problems. For the medical sector as well as the mobility sector it offers a number of valuable implementations that need to be arranged in order to maximize the benefits of AI. The government's view on possible areas to exploit AI and solve new problems seems to be limited to the medical and the mobility sector. These sectors have indeed seen great successes in recent years, brought about by AI. There are, however, many other areas that could profit from AI which are not explicitly mentioned by the government.

6.3 Higher quality

AI systems base their decisions solely on facts (being objective) as opposed to humans who may be biased or emotionally attached to a problem (being subjective). The use of AI could, thus, result in greater fairness in decision making. It also increases efficiency and safety as AI systems work faster and, ideally, make less mistakes. It could also aid society to flourish.

6.3.1 Acknowledgement and measures

We found that the increase of efficiency and safety is taken up by the government (Federal German Government 2018: 32). A fitting example is the agenda to open up administrative offices for the implementation of AI systems, paired with the provision of safe and anonymous data (Federal German Government 2018: 31). Another plan is to build a strong and open digital infrastructure in the mobility sector to improve AI systems (Federal German Government 2018: 35). This stems from AI's feature to have significantly lower error rates than humans (Federal German Government 2018: 35).

The government is aware of AI's potential to evoke a more creative society. To encourage this process, it puts forward the idea to extend the "Plattform Lernende Systeme". Thereby, the platform shall become a space for open communication between the involved areas with the hope to stimulate exchanges and the spring of new ideas.

6.3.2 Conclusion and criticism

This opportunity is partly addressed in the strategy. The government shows that it is aware of the advantages AI has in terms of safety and error rates. It includes the chance to be flourishing as society, fueled by AI. The government, however, fails to pick up on the objectivity an AI could bring for (general) decision making. It also fails to come up with an agenda on how a more creative society would look like when AI has taken over many jobs that are handled by humans today.

6.4 Scientific advancement

Insights on AI foster discussions about new ethics and stimulate research in other sciences. By coding AI, we are decoding our own intelligence and thus gain a greater understanding of human cognition.

6.4.1 Acknowledgement and measures

The importance of key technologies is recognised by the government. It mentions biotechnology as a possible connective for AI to obtain maximum potential in both areas (Federal German Government 2018: 15,16). Another structure where key technologies are located in, is the "Mittelstand". The government plans to give direct support to companies residing in this sector (Federal German Government 2018: 12). The government identifies the chance to include sociological discussions on ethics, not only for AI itself but for other sciences/technologies, too. It wants to deploy the data ethics commission as well as the enquete commission (Federal German Government 2018: 42). The chance to gain more knowledge on the brain and brain processes by studying the development of AI, is recognized by the government. It specifically mentions the field of neuroinformatics which extracts its inspiration from biological neural networks (Federal German Government 2018: 17).

6.4.2 Conclusion and criticism

The government acknowledges this opportunity and underlines the profound interactions AI has with other sciences/technologies. Similarly, it introduces novel ethical discussions that AI brings about. There are barely any concrete suggestions on how to support the intertwining of AI with other technologies to be found in the strategy. Further, the government does not explain in detail how more knowledge on human cognition may be discovered.

7 Threats

With the opportunities that AI presents, comes a plethora of threats that need to be dealt with to ensure AI fulfills its huge potential and strengthens society instead of weakening it. As AI grows more sophisticated and ubiquitous, it should be the government's main focus to address possible threats and provide measures that prevent it from causing damage. In the scope of this seminar, we identified the four most pressing threats that AI could pose.

This analysis aims to determine whether or not the government did an adequate job addressing the issues raised and offering appropriate solutions.

7.1 Loss of control

Due to the rapid pace of progress, laws and regulations are struggling to catch up. This raises the concern that the systems developed will become increasingly untrustworthy and intransparent. Considering the huge potential of AI applications, it is vitally important to understand how they work. This urgency is amplified by the private sector's pursuit of profit. Furthermore, once this technology is implemented in most of our applications and daily lives, our society is at risk of becoming overly dependent on such technologies.

7.1.1 Acknowledgement and Measures

To prevent a loss of control over systems created, the government offers suggestions like the research on the explainability of algorithms and deriving necessary regulations accordingly (Federal German Government 2018: 16). They also want to the introduce laws that place special emphasis of the individual citizen's interests and needs (Federal German Government 2018: 8). The government seeks to mitigate the threat of corporations taking advantage of the intransparency of algorithms in order to monetize them, by implementing strict regulation and oversight starting at the early stage of development (Federal German Government 2018: 19).

7.1.2 Conclusion and criticism

Although the strategy acknowledges and proposes adequate measures to ensure more transparency in the algorithms that will be developed, it does not address the risk of becoming dependent on such systems. It would have been important to mention this issue, since a dependent society makes itself vulnerable to exploitation.

7.2 Amplification of negative tendencies

The ethicality of a technology always depends on the people using it. AI has the potential to adversely affect privacy and security and enable mass surveillance that could quickly turn to social oppression. In addition, it could lead to powerful entities further consolidating their power. Countries and groups lagging the ability to utilize AI would be at a distinct disadvantage. Another significant aspect is that there is a potential of biased and discriminatory AI systems amplifying socioeconomic inequality. These biases could, for example, be caused either by distorted data or the developer's implicit prejudices.

7.2.1 Acknowledgement and Measures

An enormous emphasis is placed on securing privacy and data sovereignty. The government is conscious of the various data related issues AI could cause (Federal German Government 2018: 41). It prioritises to develop a strong ethical and legal framework that regulates how AI can be created and utilized (Federal German Government 2018: 9). Furthermore, it aims for the implementation of an observatory, a public body that has insights and monitors all the important processes, developments and key players in the field. This would enable supervision and strengthen the government's ability to react to socioeconomic changes. Aside from close observation, it wants to tackle the issue of discrimination by ensuring the transparency, traceably and controllability of algorithms (Federal German Government 2018: 38). The possible amplification of socioeconomic inequality is addressed. The brain drain caused by wealthy countries poaching experts from developing countries, combined with the prospect of massive job losses in those countries sparked by AI-driven automation could lead to an escalation of global imbalance (Federal German Government 2018: 26).

7.2.2 Conclusion and criticism

The government gives high importance to protecting privacy and data sovereignty on an individual level, but seems to neglect the threat posed by large sets of anonymous data. Information obtained by processing large amounts of such data can give insight into a population's behavioral patterns and could be used for manipulation on a societal scale. However, the variety of actions the government proposes could be a good framework to prevent manipulation and surveillance on a personal level. The success of those measures will be determined by the pace with which they are implemented. Such proceedings require swift action, since AI has already been created without regulations for a while and the rate of its progress will further accelerate in the years to come.Finally, it is important to note that surveillance and manipulation are always addressed as if they were risks only posed by external threats. It would have been important for the government to acknowledge that it could use such technologies for its own purposes of control. A clear acknowledgement of this fact, paired with a strong commitment to not take advantage of it, could have led to an increased trust of the public in the government's handling of AI.

7.3 Destruction

It's no longer a matter of whether or not AI will replace certain types of jobs, but to what extent this will happen. Its ability to automate even cognitive tasks could result in a loss of millions of jobs. Additionally, the massive amount of energy needed to power the hardware on which such complex algorithms are run, could lead to a drastic increase in resource consumption. Furthermore, by implementing algorithms into increasingly sensitive parts of society, humans are exposing themselves to malicious hackers.

7.3.1 Acknowledgement and Measures

The strategy recognizes AI-driven job loss as one of the most pressing issues society will be facing in the near future (Federal German Government 2018: 25) and acknowledges the importance of prioritising human needs in order to reap AI´s full innovative and productivity enhancing benefits (Federal German Government 2018: 25). Aside from closely monitoring the labour market, the government's main solution is centered around re-training people whose jobs are in danger of being automated (Federal German Government 2018: 25,26). To minimize the vulnerability to hackers, the government wants to ensure a high level of digital security (Federal German Government 2018: 8).

7.3.2 Conclusion and criticism

Even though the threat of job losses seems to be clear, it remains questionable if the actions proposed will prove expansive enough to be a meaningful long-term solution. The government fails to take into account more far-reaching and holistic solutions for an issue that has the potential of disrupting the labour market and society as a whole. It would have been preferable if the government had also explored more long term, societal solutions like UBI. The issue of increased resource consumption is not addressed. The reason for that might stem from the government's belief that AI is going to play a major role in solving the climate crisis. Even though there may be a substantial increase in consumption, the government seems to have the opinion that factors like the sophistication of our energy grid or smart homes will lead to a reduction in demand that outweighs the additional energy needed to run such systems.

7.4 No acceptance

Due to AI‘s complex nature and the way it's often portrayed in the media, there is a huge risk of people rejecting its adoption. Lack of understanding can lead to fear. Furthermore, widespread implementation of AI paired with a lack of acceptance by society could corrode the public's trust in our economic and political systems.

7.4.1 Acknowledgement and Measures

The government is well aware of of the public's lacking ability to have an informed opinion on the topic due to insufficient knowledge (Federal German Government 2018: 4). The importance of public understanding and acceptance of AI is discussed various times throughout the paper (Federal German Government 2018: 16). The government stresses that public acceptance relies on transparency and aims to avoid innovation being delayed or blocked due to resistance within the population (Federal German Government 2018: 43). It seeks to resolve the matter by fostering a public discourse on the topic and ensuring that the population is sufficiently educated. The offered solutions encompass a wide range of measures ranging from teaching fundamentals at an early age (Federal German Government 2018: 30) to creating a fund whose sole purpose is to educate society on the topic and enable citizens to partake in the discourse in an informed manner. How different the public reactions towards new AI-Systems are can be seen in chapter twenty nine (”How is AI-related field research conducted by police received by Twitter users?”). This chapter elaborates on various perceptions of face recognition in Germany.

7.4.2 Conclusion and criticism

Resolving this matter seems to be a priority. As such, it is extensively addressed and the measures proposed are diverse, such that if implemented timely, they could create a good framework for widespread acceptance.

8 Fields of application

Another factor that could affect the discourse are the fields of application of AI. The ontology offers ten such fields which promise to be areas in which AI could bring about innovations and useful improvements. We looked at each of these fields and searched the strategy for instances of them.

It became apparent that the government's clear focus lies on the mobility and healthcare/medical sectors. Often, the government refers to autonomous driving as one of the key technologies of today. The realization and development of autonomous driving could be an important milestone to ensure Germany's competitive position among other nations. Apart from a general agenda to tackle the healthcare sector, the strategy contains very specific examples of implementations and addresses the importance of data sovereignty. If managed well, AI could enhance this sector immensely by, for example, servicing elderly people with robots or giving emotional support to people in need. For more information on the application of AI in the medical sector, please see chapter eight (”AI in the Healthcare System: Expectation vs. Reality of Breast Cancer Detection”), chapter nineteen (”Should robots take care of the elderly? Comparing ethical guidelines to real life experiences”) or chapter twenty one (”A Modern God Complex - Doctor Who? An analysis of (un-)specialized German news articles on AI in medical diagnostics”).

The government shows that it is more or less acquainted with all remaining fields of application which include service, science, education, security and human resources. The only two exceptions are the commercial sector and the field of Natural Language Processing. The former is not mentioned in the strategy. The latter is named, but not elaborated on further.

The field of security is addressed in the strategy. Unfortunately, an aspect that is not brought up in this regard, is the potential danger of the government developing ambitions to use surveillance data for control purposes (see paragraph “Amplification of negative tendencies”).

The government broaches the production sector. However, given the considerable potential of this field of application, the government's appreciation falls rather short. There is only one mention of it in the strategy and no plans as to how production could benefit from the development of suitable AI systems. Linking this to the opportunity of AI carrying out repetitive tasks would have been feasible.

8.1 Proposal of actions

We compared the demands of Actions defined by our course to the ambitions mentioned and implied by the government. Some parts of the ontology seem to be the basis of reasoning for this paper. By supporting and investing in advancement and research in the field of AI as Germany and Europe, a subsequent effect (and most likely goal) is to counteract the technological power concentration of industrial powers such as the USA and China in the field of Artificial Intelligence (Federal German Government 2018: 6). The demands of discussing ethical guidelines, legal AI regulations (Federal German Government 2018: 38,39) and the education of the public (Federal German Government 2018: 42) call intend to launch a social and economic restructuring based on a broad public discourse (Federal German Government 2018: 43,44,45). The only measure not mentioned is a possible moratorium probably due to the fact that this would collide with the value standards intended to keep up by the German government. For more information on this, please see chapters six ("Does AI in public discourse change with different political and socio-economic systems? An analysis of the AI debate in newspapers in the emergent AI Superpowers: USA, China and Germany") or thirteen ("A comparison of different governmental approaches to prepare the public for the age of AI") of this book.

9 Public Discourse

Given the significance of a healthy discourse, we searched the strategy for specific mentions of the word "discourse". In the following, we have summarized the occurrences of it and other closely related terms. We will finalize this section with a brief conclusion of what we draw from our findings.

The government is aware of the nature of the progress of AI technologies and describes it as quick and multi faceted. The government detects the uncertain, profound and imminent changes that this progress might entail for the individual, society as a whole and our democracy (Federal German Government 2018: 44). To counteract these consequences, the government suggests various initiatives. Above all stands the motivation to earn and maintain the public's trust in AI. It professes that this requires a dynamic and constant discourse on AI related topics (Federal German Government 2018: 44). Media shall play a vital role. Here, knowledge can be shared and reviewed respectively and constructive discussions shall be held. The obligation to inform the public must likewise be assumed by educational institutions (Federal German Government 2018: 28,10). The government emphasizes the value of an explainable and transparent AI that is presented with secure and open data for it to gain the trust of the user (Federal German Government 2018: 16). Another critical obligation lies with the regulatory agencies. Norms and standardization ensure user friendliness and comparability leading to a higher quality of AI products (Federal German Government 2018: 41).

The government anticipates an instance of an imbalanced discourse in the case of an uneven acceleration of AI in the three key areas. The concern is that AI technologies in the economy see faster development yielding a dependency on AI deploying companies. For that reason, the government wants to support the scientific community and the public, so that these areas are able to give independent and competent input (Federal German Government 2018: 44).

The government is aware that chances and risks need to be communicated and discussed to the same extent. At the moment, AI is perceived controversially in Germany (Federal German Government 2018: 45). The government wants to alter this perception because the lack of knowledge and mistrust could cause constraints on innovation and development (Federal German Government 2018: 43). Ultimately, it wants society to be able to critically reflect on the topic and be informed about all its ramifications (Federal German Government 2018: 44). The government's aim is not only to get AI to be accepted in the public, but to establish an active participation of each individual in the discourse and enable the shaping of our society by AI Federal German Government 2018: 43. In this regard, the government mentions a responsibility that our generation has toward future generations (Federal German Government 2018: 44).

10 Conclusion

While the quantitative approaches at first enabled us to gain some insight into the paper and build the foundation for further analysis and interpretation, the research questions could not be answered this way alone. In the following conclusion, we therefore focused on the findings from our qualitative analysis.

To begin with, the German government demonstrates a sound awareness of the strengths of AI. The strategy is focused on the complexity and ability of AI, by explicitly naming several intellectual, complex or repetitive tasks, which can be performed by AI. As scalability and consistency are often properties of AI systems, the government implicitly covers all strengths of today's AI adequately, laying the foundation to utilize its opportunities.

By looking at the fields of research and concrete examples of where the government is looking to deploy AI, we can deduce that it is fully aware of the key strengths it has to offer. Although not explicitly stated, most of those tasks require AI to possess the four key strengths our ontology identified in order to tackle them successfully.

The government also demonstrates an encompassing awareness of the weaknesses and difficulties of AI. Some facts, such as the required acceptance by society, high expert knowledge, and enormous amount of training data are attributed a much greater value than others. The potentially high cost and some limitations of AI are mentioned only partly or are missing completely.

The government recognizes the majority of the opportunities that arise from AI technologies. It furnishes appropriate examples of each opportunity to the reader, laying the foundation for a beneficial application of AI in the next decade. There are, however, some remarks to be done. The government fails to construct its agenda in a more active way. Rather, the proposed measures seem to originate purely from a reactive mentality. The view of the government on certain implementations of AI appears short-sighted and not comprehensive enough. The latter might be rooted in formality constraints. An aspect, that is not considered at all by the government, is the opportunity to establish more objectivity for decision making processes using AI.

We can conclude that the government is aware of the most looming threats we identified in our ontology. Even though it offers an abundance of possible solutions and measures, their effectiveness will be highly dependent on the decisiveness and pace with which they are introduced. The question remains if some of the actions proposed will prove extensive enough to be a meaningful long-term solution or if more holistic approaches will be necessary to ensure sustainability and fewer unintended consequences.

Considering the above, our findings suggest that the government is generally aware of and well-informed about the possibilities and risks of AI. This not only stems from the specifically mentioned parts of the conducted SWOT-analysis, but also from the strived measures. The proposed actions for the different "Handlungsfelder" indicate an encompassing strategy covering a variety of possible problems.

In some cases, however, the government fails to take into account more far-reaching and holistic solutions. It remains questionable if the proposed actions will prove meaningful in the long-term. A few of the measures, such as the plan to put into place an observatory to check and review trends and consequences of AI, seem rather reactive and conservative. This could prove fatal since AI has already been designed without regulations for a while and the rate of its progress will only accelerate in the years to come. Swift and dynamic action is required. The success of the strategy will be determined by the pace the measures are implemented with.

It would also have been important for the government to acknowledge that it could deploy such technologies for its own purposes of control and surveillance. A clear acknowledgement of this fact, paired with a strong commitment to not take advantage of it, could have led to an increased trust of the public in the government's handling of AI.

To finalize, the government picks up on the public discourse and shows that it is mindful of its significance. However, the fact that the government states the potential mistrust of the public in AI as an obstacle for innovation and development, makes it difficult to seize the true motives of the government. The question remains if these stem from the intrinsic motivation to disseminate a well educated and mature opinion among the public or if they mainly originate from economical reasons.

11 Outlook

As this analysis does not claim to be complete we would like to propose the following approaches to get an even clearer picture of the strategy and the intentions behind it. We concluded that, overall, the measures the German government proposes are adequate and necessary. An important next step, thus, would be to not only evaluate but check if the measures are or have been turned into actions (a brief look into the national finances and the intermediate results suggests that the planned measures have not been executed yet (2020)). The intermediate results of the government include an own evaluation of its progress which would be equally valuable to look at. Another promising approach would be the analysis of the criticism papers. These represent the first instance of the discourse fostered by the German government (published on the ki-strategie-deutschland website). The criticism papers respond to each of the twelve "Handlungsfelder" and collect opinions of a wide variety of actors with different backgrounds. An Analysis of these would be an important step to explore ignored or underrepresented measures.

References

Deutsche Bundesregierung. 2020. *Nationale KI-Strategie: Fortschreibung 2020*. Deutsche Bundesregierung. Berlin.

Eduard Langwieser, Florian Fischer. 2021. *Bedeutung von Made in Germany*. https://www.german-ma.de/bedeutung-von-made-in-germany (28 March, 2021).

Federal German Government. 2018. *Artificial Intelligence Strategy*. Federal German Government. Berlin.

Merkur.de. 2018. *Merkel im chinesischen Zentrum der künstlichen Intelligenz*. https://www.merkur.de/politik/merkel-beendet-china-besuch-in-innovations-hochburg-shenzhen-zr-9897803.html (25 May, 2018).

Produktion.de. 2018. *Kanzlerin Merkel besucht Chinas Innovationshochburg*. https://https://www.produktion.de/wirtschaft/kanzlerin-merkel-besucht-chinas-innovationshochburg-210.html (25 May, 2018).

Chapter 22

A comparison of different governmental approaches to prepare the public for the age of AI

Jara Herwig, Lina Lazik, Sönke Lülf & Elisa Palme

In our modern times with rapidly changing and developing technological possibilities AI gains more and more importance in the public, private and economical life of people. Since governments play a key role in answering questions such as the country's perception, usage, legality or security, plans on how to deal with possibilities and threats must be taken into account from an official perspective. In this review the official aims and focuses of four countries, the US, Australia, Japan and Finland, were compared with respect to different fields of application. We found that some topics like health care are important to all compared countries, while other topics like firefighting are only relevant for one or two countries.

Keywords: AI | Government | Public Discourse| US | Australia | Finland | Japan

1 Introduction

Governments have an important task in shaping the public discourse surrounding artificial intelligence (AI). Since they are a driving factor in shaping the technological advancement of their country and AI appears to be a key factor in data-related industries, governments need to play a major role in creating opportunities. Strategies to do so include providing the needed education and an increase in the accessibility of smart systems.

The present article analyses steps four different governments want to take in the future or that have already been taken. To do so we present papers and official documents published by four different countries from four different continents. In order to have comparable sources, we used one main publication for each country released by a central organ of the respective government.

- For Japan there is the “Artificial Intelligence Technology Strategy” published in 2017 by the Strategic Council for AI Technology (Strategic Council for AI Technology 2017).
- “Finland’s Age of Artificial Intelligence” was published by the Finnish Ministry of Economic Affairs and Employment, also in 2017 (Steering Group of the Artificial Intelligence Programme 2017).
- In October of 2016, the National Science and Technology Council of the Obama administration published the report “Preparing for the Future of Artificial Intelligence” (Holderen et al. 2016).
- The Australian Government Department of Industry Innovation and Science commissioned an “Artificial Intelligence Roadmap” that was published by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in 2019 (Hajkowicz et al. 2019).

To further increase comparability we used the fields of applications from the shared ontology as the most important tags while leaving room for individual phenomena caused by special cultural or geological conditions. Based on the tags we created a tabMle in order to see which topics were present in all publications and whether or not topics were missing in the publication of certain governments. Afterwards cells with a lot of information were further divided regarding the government’s view on whether an aspect is seen as a strength, weakness, opportunity or threat (SWOT).

2 Definitions of AI

2.1 Finland

In their report, the Finnish Ministry of Economic Affairs and Employment is aware of the fact that defining AI can be quite a challenge. Because the report focuses on the applications of AI they choose a practical definition that fits the scope of the report. Therefore they define AI as anything, including but not limited to software and devices, that behaves in a human fashion as a result of learning (Steering Group of the Artificial Intelligence Programme 2017: 15).

2.2 USA

Also the National Science and Technology Council from 2016 is aware of the different possibilities of defining AI. Finding a definition in a regulatory framework further provides a challenge because a problem might be considered to require AI, but once it is solved it becomes a routine task. For these reasons the scope of the report includes everything that has the automation or imitation of intelligent behaviour as a goal (Holderen et al. 2016: 6,7).

2.3 Australia

The CSIRO looks at the problem of defining AI from a practical standpoint rather than from a technical or philosophical standpoint and defines Ai as "A collection of interrelated technologies used to solve problems autonomously and perform tasks to achieve defined objectives without explicit guidance from a human being." (Hajkowicz et al. 2019: 15).

2.4 Japan

The "Artificial Intelligence Technology Strategy" published by the Strategic Council for AI Technology does not provide an explicit definition of AI. From the scope of the report however it becomes clear that AI is seen as being closely related to data handling and should provide a service for humans (Strategic Council for AI Technology 2017: 4).

3 Data Collection

3.1 Finland

In the age of AI people face AI in all stages of their lives and therefore need to be able to proactively control data stored (Steering Group of the Artificial Intelligence Programme 2017: 35,36). Finland therefore presses on the need for effective data utilisation in all sectors also focussing on ethical questions (see 11). Further, the application of data modelled after MyData activities was promoted. "The purpose of MyData Global is to empower individuals [worldwide] by improving their right to self-determination regarding their personal data" (Global 2017), was founded in 2012 in Finland and acts internationally. Another project is AuroraAI, a national AI program by the Ministry of Finance, aiming to develop "an operating model for arranging public administration activities to support people in different life situations and life events so that services provided by organizations function seamlessly between different sectors". It is planned to be available in 2022 (Fourtané 2020).

The IHAN project builds a European data economy model. By combining a human-centred approach, trust, a new sense of community and the principles of sustainable growth it aims for a human-driven European data market, where companies use data responsibly and open-mindedly succeed with smart services (Luoma-Kyyny & Suokas 2018).

3.2 Australia

In order to be able to develop and be able to use AI systems on a higher level, detailed and diverse datasets are needed. For this Australia already has an open data initiative to share and use reliable data sources. This initiative releases non sensitive data as open by default and thousands of government datasets are available to the public.

3.3 Japan

Japan aims to get the best use of AI and deep learning making it necessary to feed the algorithms with information. Therefore the first step is to collect as much data as possible with the help of IoT (Internet of Things). IoT is a kind of network that collects data from every type of device, this device can be a smartphone or a sensor on a traffic light. Further, IoT analyses all the shared data and converts it in a new AI to make data more accessible, structured and usable (edureka! 2018). This collection of data can be used to analyse medical papers better. To achieve this goal they need as much data as possible digitized; in the main paper (Strategic Council for AI Technology 2017: 2) the authors criticise that Japan still has cases with not digitized data. Additionally, the government plans to make it more appealing for companies to share their data, especially private companies need to take part in the distribution of the data. Japan's focus to improve AI is advanced machine and deep learning.

3.4 Comparison

Due to the fact that AI has a great need for data, all publications mention the collection of data. However, the reports vary in terms of explicitness. Finland, Australia and Japan elaborate on the topic, while the USA discusses it more implicitly. This is due to structural decisions made by the authors and not due to the content. Data collection is mainly seen as a challenge, but all the compared countries are positive that they can adapt to it. Regardless, the context of data collection can also be seen as a threat (see: 13).

4 Health Care

4.1 Japan

One of the main reasons for collecting as much data as possible with IoT (Internet of Things) is Japan's plan to improve and focus especially on medical care. Over 40% of the country's population will be elderly by 2030 (Strategic Council for AI Technology 2017). To play a leading role in the medical wealth sector they want to develop medicine based on AI. AI could help to accumulate experiences of many researchers through deep learning. An advantage of robots and AI is that robots do not only depend on their own results, they can copy results from other robots as well, and "learn" thereby even faster (Cabinet Public Relations Office, Cabinet Secretariat n.d.). Like many other countries Japan plans to use robots and AI at nursing-care facilities to support the workers and patients, and to increase the independence of the patients. Another goal would be to lower the social security contributions. They want to achieve high living standards and high medical advancement at low cost (Strategic Council for AI Technology 2017, NewsPicks Brand design n.d.).

4.2 USA

To improve the lives of Americans AI will be used in health care. It is expected that doctors supported by AI are able to improve the treatment of patients while simultaneously lowering the costs for medical treatment. In fact, a pilot program conducted with Veterans has shown that the AI assisted treatment of combat wounds sped up the healing process and lowered the treatment costs (Holderen et al. 2016: 13). Other pilot programs are conducted to improve the health care and social situation of many Americans. One of those programs tries to fight homelessness and another aims at predicting and preventing school dropouts (Holderen et al. 2016: 14).

4.3 Finland

The Finnish government believes that chances in the public sector can be found in elderly care. Since societies get older health care costs increase. AI can be used to develop new solutions or to improve efficiency of healthcare processes, as by supporting doctors and health carers (Steering Group of the Artificial Intelligence Programme 2017: 25). The MyData project (see 3) has a secondary use of social welfare and healthcare data. Also other data networks have started to try data in test environments, in this area.

4.4 Australia

The field of AI should be further expanded in Australia, therefore it is planned to use AI in the healthcare sector. In this field robotics should be used for surgical applications to improve precision and efficacy. In addition, machine learning will be used for an earlier and more accurate diagnosis of diseases like cancer. Some AI systems even already outperform experts in the diagnosis of skin cancer using a deep convolutional neural network. Also, elderly people might get supported by AI systems in the future. Sensory Systems can monitor and assist them such that they get the possibility to live at home longer. Even for mental health issues AI might be used. A case study showed that by using an AI system and simple computer games mental disorders can be diagnosed (Hajkowicz et al. 2019: 34,35).

4.5 Comparison

Health care is an important topic for all compared countries. This can be explained by the increase of accuracy in image classification in recent years and its possible application in image based diagnostics. Another factor might be the skilled labour shortage in the sector of elderly care. Especially Japan sees an opportunity here because AI can be used to counteract the threat of a declining population.

5 Communication Bots

5.1 Finland

Overall AI offers chances to improve communication services making responses easier and faster by specialized devices. Those are independent of time and location and perform tasks better and faster with higher quality. Necessary to that end future AI systems need to improve their understanding abilities of people's needs.

Once achieved Finland plans for real accessibility of communication bots to public administration serving business as well as citizens. Communication with AI will become as normal as natural communication and a part of normal life (Steering Group of the Artificial Intelligence Programme 2017: 35,36).

5.2 Comparison

The only publication explicitly mentioning communication bots is the one of Finland. It is likely that the topic plays a minor role in the strategy of all the other countries as well. One reason why it was mentioned in the Finnish report could be that it was published by the Ministry of Economic Affairs and Employment and communication bots have a possible application in job-seeking.

6 Mobility and Transportation

6.1 Finland

Overall robotics in combination with AI is seen as a crucial part of Finland's future, in social areas as well as in transport systems (Steering Group of the Artificial Intelligence Programme 2017: 26,27). Since they are testing autonomous transport and its supporting communications solutions Finland has already reached a solid foundation to develop smart transport technologies (Steering Group of the Artificial Intelligence Programme 2017: 25).

6.2 Japan

Concerning mobility, Japan plans to make more use of autonomous driving taxis and buses, arguing that transportation should be readily available. But they do not plan to expand the mobility only for humans, it is also planned that products should be transported quicker using unmanned-following vehicles and drones. This way products can be delivered safely to their destination, including private households. It is planned to establish an AI-based service system for humans and products. This would come in extremely handy for people living in the countryside. Especially elderly people will profit from this since the bus and taxi can pick them up directly at their house and the walk to the bus station will no longer be necessary. Additionally, this would solve the problem of the shortage of drivers. The usage of autonomous traffic would also lead to fewer accidents, and therefore to more secure travelling and Japan wants to keep everything at minimal cost, every citizen should be able to afford the new transportation (Strategic Council for AI Technology 2017: 7)(NewsPicks Brand design n.d., Prime Minister's Office of Japan 2018). Japan agreed to the EU-Japan Economic Partnership Agreement, which means that autonomous cars will be constructed under the EU standards. This gives Japan the possibility to buy their vehicles from EU car manufacturers (García 2019: 79).

In addition, all traffic should be eco-friendly. The traffic will be safer with autonomous transportation, but the roads also need to be in a good state, to ensure safe transportation even further. Since the inspection and maintenance of public infrastructure is a high financial cost for Japan, it is planned to use sensors, robots and AI to inspect and maintain roads, bridges, tunnels and dams (NewsPicks Brand design n.d.).

6.3 USA

In order to improve public life, the US sees AI systems as a possibility to smarter manage traffic. A field study from 2013 showed that using smarter traffic management significantly reduced the travel time, reduced the total number of stops, the wait time and emissions in a certain area. During the rush hour in the afternoon the travel time was reduced by 29% and emissions were reduced by 25% (Smith n.d.). Other ways of improving traffic while decreasing emissions is by using autonomous means of transportation. Scientists and engineers from the National Oceanic and Atmospheric Administration used autonomous watercraft with sensors to collect data about the ecosystem along the arctic ice cap that would have been dangerous and expensive to collect by a crewed vessel (Markoff 2016). Parts of the public already made direct contact with AI-based systems in the field of driver assist features such as lane-keeping and self-parking. This potentially leads to safer traffic and eventually increases the mobility of the elderly and people with disabilities (Holderen et al. 2016: 18).

To ensure and further increase the safety of self-driving cars, several test beds have been established. Projects like the self-driving shuttles in Columbus, Ohio are a great opportunity to collect data and introduce the public to future technology (Holderen et al. 2016: 20)(Columbus n.d.).

6.4 Australia

Chances to use AI in infrastructure are also noticed in Australia. By improving efficiency and safety of transportation AI systems such as autonomous emergency braking and forward collision warning could lead to safer transportation by reducing road accidents.

6.5 Comparison

Without exception, all reports mention mobility and transportation. This is not surprising because autonomous vehicles already play a major role in the public discourse surrounding AI. When the topic is discussed the strength of autonomous systems like reduced emissions, increased safety and increased availability and flexibility outweigh their weaknesses like ethical dilemmas.

7 Education

7.1 Australia

AI will change the job market and work of many people. Workers will need to train and reskill to catch up with the changes that AI brings. To make this easier Australia plans to start early with integrating useful skills in the education system (Hajkowicz et al. 2019: 48). Of importance is also that people build trust in AI since some of the human control is handed over to AI. Trust needs to be built for two things in particular, for the technology itself and for the company behind the technology. To build this trust technical explanation and transparency is of need. This means that the work and record of the company must be transparent as well as the workings of the technology.

7.2 USA

The US government sees itself in a key role to create a workforce that lives up to the challenge AI is going to bring. Education in the field of data science should be provided as early as possible and therefore start in primary schools. The initiative 'Computer Science for All' should provide every young American with the basic skills needed for becoming creators in a technology-driven future, as a key part of STEM (science, technology, engineering, and mathematics) education (Holderen et al. 2016: 26,27). Early education is especially important when facing the challenge of making the STEM workforce more diverse with respect to both gender and race (Holderen et al. 2016: 28). Furthermore, Colleges play an important role for workers that are expanding their skills and unemployed people that are trying to return into the workforce (Holderen et al. 2016: 27). In addition to learning about AI or data science every student should be exposed to training in and discussions about ethics. This should help these practitioners to understand their responsibilities and put good intentions into practice (Holderen et al. 2016: 32).

7.3 Japan

If Japan wants to give AI a big role in everyday life, they also need to prepare society for the next state. And to make society more sensible for AI Japan's government plans to reform the education system, including preschool as well as universities. To give the children a better feeling for robots and AI programming should be already taught in preschool and at universities, the focus is expected to be more on AI. Japan wants to improve the collaboration for research between universities and companies (Strategic Council for AI Technology 2017: 10). The prime minister even induced the "Artificial Intelligence Technology Strategy Council" (García 2019: 28). For workers, they plan an education program, so everyone improves knowledge about parts of AI which are important for their sectors.

7.4 Finland

Education will play an important role in Finland's smart technology revolution. Priorities in education need to be changed towards the application of AI and its effects, rather than the technological and mathematical background behind it (Steering Group of the Artificial Intelligence Programme 2017: 39). Aiming to ensure a more rapid and easy adaptation of AI in early 2018 the Finnish AI Business Program was launched to spread more AI knowledge outside of larger technical growth centers. Further the Finnish Centre of AI of the Aalto University in Helsinki and VTT, one of Europe's leading research institutions with the aim to "help companies and society in solving global challenges by utilising science and technology" (of Finland Ltd. 2020), promoted AI research, its use and application and a self-assessment tool for companies, or AI index, was published (Steering group and secretariat of the Artificial Intelligence Programme 2019: 63-71).

Additionally, Finland hopes to get top-level expertise and attract top experts. Therefore the competence needs for AI and its education training were analysed and the first steps for the development of new education and training were taken, as the free online course ”Elements of Ai” by the University of Helsinki. Finland's attraction for specialists was increased by Finnish Centre of AI (FCAI), aiming to attract talents nationally and internationally. At universities of applied sciences a Master of AI degree was promoted and training methods for companies were established (Steering group and secretariat of the Artificial Intelligence Programme 2019: 72-78).

7.5 Comparison

The easiest and arguably most important sector for advancing in the field of AI is education. Education in data science and AI does not only bear the needed high level experts but also increases the discourse surrounding the topic and helps integrate the principles and implications of the coming change into the culture of each nation. Unsurprisingly, all reports are positive towards AI education, either stating a well-functioning education system as a strength or finding opportunity in future reforms.

8 Society

8.1 Finland

From the Finish perspective, the social question is concerned with limits of technology in citizens and organisations activities. Due to unknown factors such as the pace of technological development and its actual influence on society the exact impact of AI is however hard to predict (Steering Group of the Artificial Intelligence Programme 2017: 36,37). Overall it can be predicted that new and smart technologies will transform the whole society, in Finland as well as globally (Steering Group of the Artificial Intelligence Programme 2017: 32).

A society's reaction to AI depends on active and passive elements. Passively societal institutions filter the effects of technology on practical working life and society, actively proactive social regulatory systems guide solutions with technological development into desired direction, focusing on the question of what a good AI society is (Steering Group of the Artificial Intelligence Programme 2017: 39,40).

Focusing on the changes in the nature of work by AI in 2018 a review on work in the age of artificial intelligence was published. It deals with the effects of AI on trends in economy and employment, the changes of the work and labour market, the necessary reforms in education and skill maintenance and the ethics behind AI in working environments. In general, responsibility is seen on the sides of employers, employees and the society itself to update their own skills and to create a safe and fair demand-based market for work, education and training (Steering group and secretariat of the Artificial Intelligence Programme 2019: 97-101) (Koski & Husso 2018: 5). It is likely for the amount of medium-salary jobs to decrease, while the amount of low and high-salary jobs increases. Structural changes, as duty takeovers of doctors or lawyers by AI, are likely. Jobs requiring personal contribution, flexibility, creativity and skills in problem-solving, presentation, communication or interpersonal tasks will become more important and the productivity of less educated people may increase causing an overall more equal society (Steering Group of the Artificial Intelligence Programme 2017: 37,38).

Aiming to build the world's best public services using the AuroraAI program preliminary studies have been carried out to create the foundations of a society in the age of AI. The key aspect is to do so in a human-centred, ethically sustainable way. The first AuroraAI trial enabled mutual interactions of smart applications and public services. Subsequently, essential life and business-based events were identified, to form a human-centred service ecosystem and to set up a support team to change operations (Steering group and secretariat of the Artificial Intelligence Programme 2019: 84-89).

8.2 Japan

With the help of AI, Japan wants to achieve one of its biggest goals: society 5.0. Currently, the society is in the 4.0 state, the "Information Society". They describe the 5.0 state as "a technology-based human-centered society" (Prime Minister's Office of Japan 2018). While the population becomes older and older, the birth rate is very low. A problem the government wants to solve by AI as well. They plan to fix the problem by creating human-matchmaking services (e.g. dating apps) which are analysed by AI ("Japan to fund AI matchmaking to boost birth rate" 2020).

8.3 Comparison

The impact that AI is going to have on society is expected to be huge. The different degrees to which this topic is discussed can therefore be surprising. However, most of the publications focus on steps that need to be taken and sectors that have to develop in order for AI to improve and not on the implications of an AI-driven change. Reaching society 5.0 is a concrete goal of Japan that already influenced some political decisions. The finnish report on the other hand sees the implications of AI as a threat, possibly leading to unemployment but at the same time as an opportunity to create a more equal society.

9 Economy

9.1 Japan

As a result of an older growing society, less Japanese citizens will be able to work, in order to fill the gaps AI needs to support the economics and replace those missing workers. The usage of robots and AI in economics might also improve productivity.

9.2 Australia

Due to the aging workforce in agriculture of Australia robotics could be an opportunity to reduce the workload of farmers and enhance productivity. By AI improved weather forecasts will be a help to adapt the farming process and watering times. In Addition to applying AI in agriculture it could be used in mining as well, for instance by using computer vision to improve safety and analyze geological data by machine learning to enhance mineral discovery (Hajkowicz et al. 2019: 40,41).

9.3 USA

The US plans to invest in long-term research. This includes considerations on predictable outcomes as well as high-risk investments that could potentially lead to high-reward payoffs. The possible long-term reward can be compared to the reward of the World Wide Web. It is important to note that investments will not only be made in the field of software but also in the field of hardware development (Biegel & Kurose 2016: 16,17).

9.4 Finland

In Finland long-time research of institutes, companies and public organisations, the supportive legislation system and the rapidly growing amount of a start-up ecosystem give a good foundation (Steering Group of the Artificial Intelligence Programme 2017: 28,29). This represents a chance because economic growth is seen as the basis for high-quality public services and a well-functioning society(Steering Group of the Artificial Intelligence Programme 2017: 35). Companies will incite the development and application of AI (Steering Group of the Artificial Intelligence Programme 2017: 32). Technological development is seen as the most significant factor in economic growth for both private and public sectors, however increasing AI can also be seen as a threat in economy with respect to amount and quality of working. In the future AI will become a support for humans and thereby increase the quality of the work done (Steering Group of the Artificial Intelligence Programme 2017: 18f,19). Overall, production costs will decrease and at the same time allow workers to have more time for interactions which can only be done socially (Steering Group of the Artificial Intelligence Programme 2017: 24). It can be assumed that the working world has to face two phases. First new technologies related to products, services and production processes may cause unemployment, second new technologies, business-driven organisational and social innovations may create new jobs and career opportunities (Steering Group of the Artificial Intelligence Programme 2017: 36).

Finland had the goal to enhance business competitiveness through use of AI. Actions already taken are the acceleration of AI, for instance by peer learning or sharing the best practise and solutions and a general support of expert assistance, important for networking and funding applications (Steering group and secretariat of the Artificial Intelligence Programme 2019: 47-52).

They also want to make bold decisions and investments. The capital loan funding for Growth Engines of novel platform companies using data as well as investments in the leading-edges of research, learning, data management and computer infrastructure have already been achieved (Steering group and secretariat of the Artificial Intelligence Programme 2019: 79-83).

Furthermore, Finland aims to establish new models of collaboration that address the AI program from different angles operating as a network of networks. Stakeholders from within and outside administrations have identified and solved bottlenecks of digitisation (Steering group and secretariat of the Artificial Intelligence Programme 2019: 89-93).

9.5 Comparison

When it comes to the economy every country faces different challenges and builds upon individual strengths. For Japan AI presents an opportunity because missing workers pose a threat. For Australia agriculture plays an important role in the economy while the geological situation can be challenging with respect to the use of water. For the US the silicon valley has been a financial success in the past and is likely to create more opportunities in the future. Finland too has high hopes for the future, with supporting start-ups Finland plans to build a solid foundation for the future.

10 Security

10.1 Finland

In security, AI brings benefits as well as risks. While AI can offer solutions in complex and surprising situations, the need for higher security is rising due to more uncertainty in the world and more powerful technical threats in private and economic life. Finland's focus of AI-security lays on the protection of individuals and privacy (Steering Group of the Artificial Intelligence Programme 2017:

27). Further, the national cybersecurity strategy has been improved and updated. Overall the social importance of data security was stressed and opens chances for companies investing in it and providing various services. Internationally Finland has been part in the opening of a European Cybersecurity Network and Competence Centre (Steering group and secretariat of the Artificial Intelligence Programme 2019: 109-113).

10.2 USA

In the field of cybersecurity the US poses AI as a possible threat since cyberattacks could be more sophisticated and efficient. On the other hand, AI also holds the potential to increase cybersecurity and make it more available while lowering cost, because secure systems no longer have to be developed and installed by highly trained experts (Holderen et al. 2016: 36).

10.3 Japan

Japan has identified ‘information security’ as a priority (Strategic Council for AI Technology 2017: 4). The development in this field is strongly connected to various other fields. It is considered important because it makes critical systems reliable, stable and secures the confidentiality of personal data (Strategic Council for AI Technology 2017: 8).

10.4 Australia

Australia plans on implementing smart systems for many applications. These systems are more vulnerable to cyberattacks than analogue systems that are in use. Therefore, Australia has identified a need for cybersecurity (Hajkowicz et al. 2019).

10.5 Comparison

It comes as no surprise that all countries have identified cyberattacks as a possible threat in the age of AI. Nevertheless are the reactions diverse. Japan and Australia try to counteract these threats by investing in cybersecurity and taking other measures. The US and Finland see AI itself and the change it brings as possible solutions.

11 Ethics

11.1 Australia

In general standards and ethics are of large importance for the Australian appliance of AI. To tackle this problem they created an ethics framework with eight core principles. These say that AI should generate a greater benefit than it costs, it should do no harm and comply with regulations and laws. Also privacy must be protected, fairness should be kept and the algorithms impact must be transparent, explainable and contestable.

11.2 USA

Creating AI that is in line with ethical guidelines is seen as a challenge in the US because ethical principles are often formulated in vague terms and therefore not easily translatable into an algorithm. Furthermore, autonomous systems raise a set of never before asked ethical questions. A possible solution is presented in terms of a multi-disciplinary approach to create datasets including examples that can be used as legal and moral corner cases (Biegel & Kurose 2016: 27).

11.3 Finland

New technologies always impact various aspects of society and depend on institutional and cultural regulations. By restricting ways of technological application Finland focusses in its ethical questions on topics such as the openness of health data, location monitoring or the use of robots in nursing and health care (Steering Group of the Artificial Intelligence Programme 2017: 36,37). Since the country aims to develop AI into a trust-based human-centred direction, discussions on AI ethics and the use of AI in public sectors from the view of ethical and societal acceptability have already been conducted. The legality and ethics for the AuroraAI service were defined and international ethical discussions were held. Overall the citizens' understanding on the ethical viewpoint of AI was enhanced (Steering group and secretariat of the Artificial Intelligence Programme 2019: 102-108). Further focussing on ethical questions of data collection as in MyData in December 2018 an ethical information policy report was published by the Ministry of Finance (Global 2017). This way Finland has even become one of the leaders in the ”High-Level Expert Group on Artificial Intelligence”, a group preparing for instance ethical guidelines for trustworthy AI in Europe (Steering group and secretariat of the Artificial Intelligence Programme 2019: 93-96).

11.4 Comparison

Ethical implications of AI have already played a significant role in the discourse surrounding AI. Since new technology always creates novel situations it also raises never asked ethical questions that countries have to answer individually given cultural and legal differences. Overall the paper of Japan views AI in a very positive way and as a solution to many problems. At the same time they do not critically question and mention ethical parts and dilemmas.

12 Legislative

12.1 Finland

In Finland, regulation and legislative limits of AI will be organized by legislative filters, while realistic economic aspects make clear that the economy is often slower than technology itself (Steering Group of the Artificial Intelligence Programme 2017: 36,37).

12.2 Comparison

AI raises not only ethical but also legal questions. The countries are not very explicit with regards to the legislative future when it comes to AI. This can be due to the fact that the legal implications of this technology are not fully understood yet or regulations have to be found on a regional level rather than a national one.

13 Criminal Justice

13.1 USA

Focusing on justice the US identifies opportunities of AI to be used in order to improve the criminal justice system "including crime reporting, policing, bail, sentencing, and parole decisions" (Holderen et al. 2016: 14). This opportunity however is coupled to a few risks. On one hand, there are concerns that the available data is biased and not enough, on the other hand AI based systems used in the criminal justice system need to be accountable (Holderen et al. 2016: 30).

13.2 Comparison

Criminal Justice has been the subject of many public discussions in the US. It is expected that this topic gets addressed in reports on a technology that could reform the criminal justice system. Other countries that seem to have less of a problem in this area are not expected to mention the topic in such high-level reports.

14 Electricity and Firefighting

14.1 Australia

Special to Australia is the concern of the usability of AI in fields of electricity and firefighting. In the energy sector AI could bring a more efficient use of electricity which reduces energy security concerns and power outages (Hajkowicz et al. 2019: 36,37). Another application for AI in Australia is to help firefighters fight bushfires. AI is able to map forest fire fronts and simulate fire-spread which helps firefighters to concentrate their work on the correct spots. Systems like these already exist like the CSIRO Data61 "Spark" which operates in real time(Hajkowicz et al. 2019: 12).

14.2 Comparison

Due to its geological specifics Australia has struggled the most with bushfires out of all the compared countries.It comes as no surprise that none of the other countries have identified AI as an important opportunity in this field.

15 Conclusions

15.1 Australia

Australia has a lot of plans for AI that will touch the lives of Australians in many different ways. The hope is that these plans will boost Australia's productivity and improve the overall quality of life. Consequently, Australia focuses on the opportunities provided by AI. However, the publication commissioned by the Australian government was published in 2019 there has not yet been a follow-up report.

15.2 Japan

All in all, it can be said that Japan has high hopes in AI to fix many of their problems. Besides considering short term consequences, Japan further plans on what they can do in a more distant future. It is important for Japan to quickly gain control of their older growing population, not working society and low birth rate, otherwise they will have a too small workforce in a few years which could also make research for AI more difficult. Co-operations with international partners, like the EU, are therefore a great and necessary idea.

The paper was released 3 years ago, and there have been a few changes. By now some banks operate with telephonic customer services by using speech and voice recognition. Not only banks use AI, some of Japan's railways can identify intoxication of passengers using AI as support. When it comes to job candidates AI helps (some) companies to analyse people's analytics and recruit the most fitting of all candidates (García 2019: 27). They also started to invest more money in AI-related start-ups and continue to make start-ups more appealing for private people and companies (García 2019: 23) (Strategic Council for AI Technology 2017: 11). During the corona pandemic Japan was quick with finding solutions, they had robots walking through malls and airports and were able to detect if people had a fever, isolate the person and even call a doctor if necessary. And the famous robot Pepper was assisting hotels to keep the employees and guests as safe as possible (Dirksen & Takahashi 2020: 28). By this modern possibilities of AI were merged into the everyday awareness of the country's citizens.

15.3 Finland

In their first report from 2017 Finland provided eight key actions. In 2019 a follow-up has been published (Steering group and secretariat of the Artificial Intelligence Programme 2019), in which the government reflected on the process made concerning the eight initial key actions. Three more key actions, which have occurred over the country's course of the past two years, were added.

Overall in the key actions planned and taken it gets clear that Finland sees the economy as incitement behind all societal changes. Due to developments in the economy – so the prediction – AI will gain an ever larger role in the private every day and working life of people and thereby change the way they see and interact with new technology. By educating on every level, beginning at the youngest ages and ending at elderly people, Finland hopes to create a society capable of handling the new technological threats and chances. By introducing programs such as AuroraAI they try to make data easier available for everyone, but still attempt to form a trust-based human-centred service ecosystem with security and protection of the citizens as core conditions. According to Finland in future AI will play a large role in everybody's life and therefore take a big role in the public discourse and discussion.

15.4 USA

In the governmental publications, the US focuses on the opportunities that AI will provide in the future. The government is positive that the US will keep on being one of the leading nations in the fields of AI and technology. The strategic plan published by the national science and technology council gets updated regularly. However, since this strategic plan is formulated rather vaguely and does not make concrete suggestions it is hard to pinpoint the progress already made.

The greatest opportunity can be identified in the field of education. Early education in computer science, especially in AI, can create possibilities for a future oriented workforce. Increased qualification in the field of AI will also lead to a more constructive public discourse on the topic.

16 Final conclusion

Taking everything into account, all four countries discussed a lot of topics and overall seem to have high hopes regarding the future of AI. Many topics such as health care, education, mobility and economics were talked about in all of the publications. However, some topics were specific to a single or a small subset of countries. For example Australia discussed firefighting, the US mentioned criminal justice, Japan talked about society 5.0 and Finland specifically addressed communication bots. In the overwhelming majority of cases a topic either is seen as an opportunity or the already existing structure is seen as a strength.

After reading the different governmental approaches it gets obvious that AI is a global responsibility and requires collaboration on a high level.

References

Biegel, Bryan & James Kurose. 2016. *The national artificial intelligence research and development strategic plan*. National Science & Technology Council. https://www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf.

Cabinet Public Relations Office, Cabinet Secretariat. N.d. *Ai to advance regenerative medicine*. The Government of Japan. https://www.japan.go.jp/technology/innovation/aitoadvance.html.

Columbus, Smart. N.d. *Self-driving shuttles.* https://smart.columbus.gov/projects/self-driving-shuttles.

Dirksen, Nicole & Sonoko Takahashi. 2020. *Artificial intelligence in Japan 2020.* https://www.rvo.nl/sites/default/files/2020/12/Artificial-Intelligence-in-Japan-final-IAN.pdf.

edureka! 2018. *Internet of Things (IoT) | What is IoT | How it Works | IoT Explained | Edureka.* https://www.youtube.com/watch?v=LlhmzVL5bm8.

Fourtané, Susan. 2020. *Auroraai: Finland's national artificial intelligence program.* Accessed: 2021-01-13. https://interestingengineering.com/auroraai-finlands-national-artificial-intelligence-program.

García, Guillermo. 2019. *Artificial Intelligence in Japan.* https://www.eu-japan.eu/sites/default/files/publications/docs/artificial_intelligence_in_japan_-_guillermo_garcia_-_0705.pdf.

Global, MyData. 2017. *About - mydata.org.* Accessed: 2021-01-13. https://mydata.org/about/.

Hajkowicz, Stefan, Sarvnaz Karimi, T Wark, Chen Cai, M Evans, N Rens, David Dawson, Andrew Charlton, Andrew Brennan, Moffatt Corin, Sriram Srikumar & Jun Tong. 2019. *Artificial intelligence: solving problems, growing the economy and improving our quality of life.* CSIRO Data61, Australia. https://data61.csiro.au/en/Our-Research/Our-Work/AI-Roadmap.

Holderen, John, Megan Smith & Executive Office of the President. 2016. *Preparing for the future of artificial intelligence.* National Science & Technology Council. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

Japan to fund AI matchmaking to boost birth rate. 2020. *BBC News.* https://www.bbc.com/news/world-asia-55226098.

Koski, Olli & Kai Husso. 2018. *Work in the age of artificial intelligence: four perspectives on the economy, employment, skills and ethics.* Ministry of Economic Affairs & Employment.

Luoma-Kyyny, Juhani & Jyrki Suokas. 2018. *IHAN - What is it about?* Accessed: 2020-12-20. The Finnish Innovation Fund Sitra. https://www.sitra.fi/en/projects/ihan-pilot-projects/#what-is-it-about.

Markoff, John. 2016. No sailors needed: robot sailboats scour the oceans for data. *The New York Times.* https://www.nytimes.com/2016/09/05/technology/no-sailors-needed-robot-sailboats-scour-the-oceans-for-data.html.

NewsPicks Brand design. N.d. *Realizing society 5.0.* The Government of Japan. https://www.japan.go.jp/abenomics/_userdata/abenomics/pdf/society_5.0.pdf.

of Finland Ltd., VTT Technical Research Centre. 2020. *What is VTT.* Accessed: 2021-01-14. https://www.vttresearch.com/en/about-us/what-vtt.

Prime Minister's Office of Japan. 2018. *Society 5.0: concept.* https://www.youtube.com/watch?v=SYrv6kOsU1o.

Smith, Stephen. N.d. *Smart infrastructure for urban mobility.* https://cra.org/ccc/wp-content/uploads/sites/2/2016/06/Stephen-Smith-AI-slides.pdf.

Steering group and secretariat of the Artificial Intelligence Programme. 2019. *Leading the way into the age of artificial intelligence.* Ministry of Economic Affairs & Employment.

Steering Group of the Artificial Intelligence Programme. 2017. *Finland's age of artificial intelligence.* Ministry of Economic Affairs & Employment. http://urn.fi/URN:ISBN:978-952-327-290-3.

Strategic Council for AI Technology. 2017. *Artificial intelligence technology strategy.* The original website is no longer available. https://web.archive.org/web/20200527083705/http://www.nedo.go.jp/content/100865202.pdf.

Chapter 23

Ethical guidelines in the European judicial system

Hanna Algedri & Till Holzapfel

The European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment is roughly 70 pages long and was drafted by the European Commission for the Efficiency of Justice (CEPEJ) . Initially it was a reference document for public discourse on the usage of AI in the judicial system. Since then it has become one of the most referenced documents on international political summits. Consequently, it will most likely play a key role in the development of a certificate for AI tools, that is official and legally binding for tools operating within the council of Europe's 47 member states. As a qualitative analysis, we discuss the general structure and objective of the Charter, focusing predominantly on the content than the style of the document. In addition, we put a focus on the Charters chapter about Predictive Policing as an example application area. Overall the Charter has a rather strong focus on the positive potential of AI tools, but it also outlines ethical guidelines and repeatedly emphasises that they need to be adhered to without exception in order for an AI tool to be incorporated into the judicial system.

Keywords: European Commission for Efficiency of Justice (CEPEJ) | Ethical principles for AI | Predictive Policing

1 Introduction to the ”European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment

1.1 What is the Charter and who created it?

The Charter was written by the CEPEJ [1] as commissioned by the Council of Europe Ronsin et al. 2018. The council describes itself as “the continent’s leading human rights organisation” (p.78)[2] and consists of 47 member states, including all members of the European Union. They have signed the European Convention on Human Rights, a treaty designed to protect human rights, democracy and the rule of law. The European Court of Human Rights oversees the implementation of the Convention in the member states. The Charter itself has no legally binding properties, since it was created as the first explorative step to create laws later on. It is however the most prominently used reference document in discussions about the topic globally and therefore of special interest.

In January of 2021 the CEPEJ published a feasibility study called “Possible introduction of a mechanism for certifying artificial intelligence tools and services in the sphere of justice and the judiciary”[3] that builds directly on the principles developed within the Charter. Furthermore, on the road map[4] of the CEPEJ cyberjustice work group for 2021 are plans to create an institution called ”European Cyberjustice Network (ECN)” which is described as follows:

> “The ECN should be composed of the contact points from the competent authorities within the member States having expertise in the field of cyberjustice and artificial intelligence. The Network should exchange and disseminate information on the situation and best practices and support initiatives in member States. It should also initiate proposals and enable a platform for bi-or multilateral co-operation in the field of e-justice.” (CEPEJ 2020).

1.2 Motivation behind the Charter

But why was such a Charter commissioned in the first place? According to the authors, the development of AI-justice tools is for the most part happening within

[1] the respective working group members can be found here: https://rm.coe.int/cepej-bu-2020-2-composition-gt-2020-2021/16809e2b4c

[2] any time that we quote from the Charter we will only reference the page number

[3] https://rm.coe.int/feasability-study-en-cepej-2020-15/1680a0adf4

[4] https://rm.coe.int/cyberjustice-roadmap-en-cepej-2020-14/1680a0ae12

the private sector and has insurance companies, lawyers and legal service providers as its clientele. These tools currently focus on reducing legal uncertainty and even predicting judicial decisions. The authors of the Charter are under the impression that "public decision-makers are [...] increasingly solicited" (p.14) by these private entities to integrate these tools into public policies. They also repeatedly state their opinion that these tools could have many benefits for the judicial system, mostly but not only with regards to its efficiency in data processing. The incorporation of AI into our judicial system is viewed as a step that should not be taken without a prior investigation of the possible consequences and the Charter is supposed to be the first step in this process.

So whom exactly is the Charter supposed to inform about this development? At this point the Charter enters public discourse: It is intended to inform "public and private stakeholders responsible for the design and deployment of artificial intelligence tools and services that involve the processing of judicial decisions and data" as well as "public decision-makers in charge of the legislative or regulatory framework, of the development, audit or use of such tools and services." (p.5).

> "It is essential that any public debate involves all the stakeholders, whether legal professionals, legal tech companies or scientists, to enable them to convey the full scope and possible impact of the introduction of artificial intelligence applications in judicial systems and devise the ethical framework in which they must operate. Subsequently, this debate could go beyond a pure "business" framework, **involving citizens themselves** []." (p.16)

1.3 Application of the Charter

Even though the guidelines developed within the Charter are not legally binding as of now, they are supposed to do more then spark up and inform the debate about this topic.

> "The principles of the Charter should be subject to regular application, monitoring and evaluation by public and private actors, with a view to continuous improvement of practices." (p.6).

This statement is addressed at already operating systems in the private sector as well as tools that are currently still in development. It is important to note that it does not only refer to the programs as such, but also the way in which they are developed and used. The principles mentioned here refer to the five basic ethical principles that are developed within the Charter and will be discuss in the next chapter. Each of them includes a one sentence summary on what to do as a developer of AI-justice tools in order to ensure that your product is compliant with the Charters principles.

1.4 Structure of the Charter

Now that we have an idea about the goals of the Charter and why it was commissioned, let us shortly describe how it is structured. First up, the Charter discusses the inner workings of "mass case-law data processing systems" (p.15)[5] including their technical and theoretical limitations. The Charter generally tries to encourage as much understanding of AI-justice tools as possible to enable an informed discussion regarding their application. Secondly, the Charter analyses the benefits and risks that these tools are currently perceived to have. Example benefits that are commonly put forth by supporters of these tools include an increase in "transparency, predictability and standardisation of case-law" (p.15), while critics highlight especially the issue of bias. In this context, the Charter puts forth examples that are taken to be a positive application of such tools.

> The Charter "generally advocates the use of AI by legal professionals according to their needs, provided that due regard is shown for the individual rights guaranteed by the European Convention on Human Rights (ECHR)[6] and Council of Europe standards, particularly in criminal matters." (p.16)

The explicit mentioning of criminal matters here is one of the reasons why we chose predictive policing, the AI supported prediction of where a crime might happen, as the part of the Charter which we will discuss in more detail. Another reason is that it operates on the first step of any criminal trial[7], the detection of a crime. Lastly we belief that the risks as well as benefits of such tools are easily relatable, even without any expert knowledge. They have been discussed in the news around the world, particularly in the US but also here in Germany as discussed in the chapter "How is AI-related field research conducted by police received by Twitter users?".

[5] This refers to large amounts of data on past court decisions and would generally fall under Big Data: "Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software." https://en.wikipedia.org/wiki/Big_data

[6] https://www.echr.coe.int/Pages/home.aspx?p=basictexts&c

[7] for clarification: we will refer to criminal trials / court cases simply as trials

2 The five ethical principles

The CEPEJ developed five ethical principles on the use of AI in the Judicial System.

Principle of respect for fundamental rights This principle refers to the right, that AI in Judicial Systems should be in full compliance with the fundamental rights as laid out by the European Convention on Human Rights and the Convention of the Protection of Personal Data. This principle includes the right to access a judge and the right of a fair trial. This is ensured by rules prohibiting violations of the fundamental rights. Furthermore, the judges independence in the decision-making process has to be ensured.
Corresponding advise for developers by the Charter:

"Ensure that the design and implementation of artificial intelligence tools and services are compatible with fundamental rights, including the right to protection of personal data." (p.77)

Principle of non-discrimination This principle states that stakeholders must ensure that methods do not support discrimination and do not lead to deterministic analyses through the classification of data related to individuals or groups of individuals. Nonetheless, if discrimination is identified they have to neutralise the risk. Especially sensitive data like ethnic origin, political opinions, sexual life or orientation etc. must be taken care of. In contrast, AI tools combating discrimination should be supported.

"Specifically prevent the development or intensification of any discrimination between individuals or groups of individuals" (p.77)

Principle of quality and security For the development of these tools, the designers of Machine Learning algorithms, and experts of the judicial system, law and social science should build a team to share expertise throughout the design cycles. The data which is based on judicial decisions, should be drawn from certified sources and should not be modified before the end of the learning mechanism. Furthermore, the systems integrity and intangibility must be ensured.

"With regard to the processing of judicial decisions and data, use certified sources and intangible data with models elaborated in a secure technological environment" (p.77)

Principle of transparency, impartiality and fairness There are different approaches according to balance between the need for transparency, impartiality, fairness and the intellectual integrity during the design and operating chain. Either complete and public technical transparency could be achieved or independent authorities or experts could certify and audit the processing methods while public authorities could grant the certification.

"Make data processing methods accessible and understandable, authorise external audits" (p.77)

Principle "under user control" The user autonomy should be increased by informing and including users of the different options that are available, for example the use of AI Tools during trial. Therefore, it must be possible to review the judicial decisions and the data used to produce these. Relating to the *Principle of non-discrimination* the user should be informed, that they have different options regarding the process, i.e. the right to access a court and the right to legal advice.

"Preclude a prescriptive approach and ensure that users are informed actors and in control of their choices" (p.77)

3 Criminal Matters

3.1 Before The Trial

Predictive Policing

One way to use tools before a trial is predictive policing. As the name suggests, software is used to predict who or where the next crime will be committed. The Charter cites the "No Fly List" maintained by the United States federal government's Terrorist Screening Center (TSC) as a well-known example, which collects and algorithmically evaluates data on potential terrorists. Other applications of these algorithms include for example the detection of fraud or money laundering. For everyday city police work, the first step is to detect crimes that occur with a certain regularity, such as theft or burglary. These probabilities are displayed on a map, which then marks the particularly critical locations as hot spots for police patrols. This type of predictive policing is called predictive criminal mapping. Usually the software is fed with data sets from police reports or other technologies. This obviously has the advantage of detecting crimes ahead of time or catching the perpetrator during the crime with a greater probability, but there are also considerable weaknesses. The Charter cites the emergence of a vicious cycle (see Figure 1) and self-fulfilling prophecies in this regard. Furthermore, various positive aspects of predictive policing are mentioned without coming to a final and evaluative judgement. Two examples of the benefits of predictive policing are given in the context of financial crimes and also child abuse, where the algorithm provides both a time and accuracy benefit through analysis. While the advantages of predictive policing are clearly stated and supported by examples, major weaknesses, such as self-fulfilling prophecy and also the possibility of racial profiling are not further elaborated on. The Charter as a whole is often more explicit about the positive than the negative potential of AI justice tools, which could possibly bias the readers perception. Our impression is that they might think the adherence to their ethical principles should already cover all the possible negative outcomes.

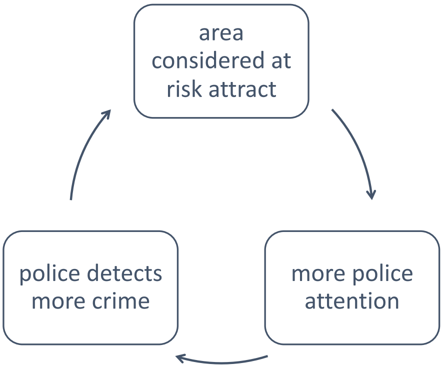


Figure 1: vicious circle of predictive policing

3.2 During the Trial

The requirements for a tool to be used during a trial are, among other things, to support the judge in making a prediction about the probability of recidivism. The Charter refers to the only tool in Europe, or rather in England, called Harm Assessment Risk Tool (HART). HART was developed with the help of the University of Cambridge and is intended to help determine whether a suspect is likely to commit a crime again. Among other factors, the Charter cites that zip code and gender play a role for the calculation. It also cites HART as a good way to identify challenges for tools used during the process. However, since HART is the only "During the Trial" tool in Europe, the Charter refers to findings from the US.

To show the challenges and also dangers of such a tool, the case of Correctional Offender Management for Profiling for Alternative Sanctions (COMPAS) software is explained. This software is used in the US to help judges assess recidivism rates, but discriminatory factors were discovered, including things like whether one owned a home phone or the family background. This resulted in discriminatory software that attributed twice as much recidivism to African-Americans as to other ethnic groups. As a result, the false positive rates were very high. The Charter explains this by saying that such software shows the social and also economic fragility of these groups. Moreover, the software was developed by a private company. This turned out to be another obstacle for the transparency of the algorithm, the code could not be reviewed and checked by an independent authority or expert. Not only does this negative example show the polarisation of injustices of a society, but also the difficulties of identifying discrimination. Discovering and fixing these in the code or algorithm could trigger a conflict with the "intellectual property".

3.3 Challenges

In this part of the Charter the previous findings are summarised in order to provide guidelines and approaches for the development and use of tools in the judicial system. In particular, it sets requirements for minimising the risk that such tools can entail, and thus for maximising their potential. In particular, the Charter points out that public decision makers and judicial stakeholders must be aware of the possible dangers and should already get involved during the development of the tools, since otherwise the personal freedom of an individual is in danger. Furthermore, the Charter refers to possible arguments of proponents, which, however, should be treated with caution and should not be weighed against the possible dangers. The effectiveness, efficiency and possible objectivity are good reasons in favour of these tools. Whereas when the criminal past or family background play a role in the evaluation, the past would decide the fate, while the future behaviour is assessed, are reasons against these tools. A human being, in this case a judge, can also include different weights in his decision, whereas to this extent it is not possible with an algorithm. The Charter also mentions that it is worthwhile to look at tools from the US for insight purposes, but not to forget the significant differences in legislation between Europe and the US.[8] In Europe,

[8] One example would be the COMPAS tool we mentioned in the last section (see 3.2). In the US the intellectual property of the code is an acceptable reason for not offering technical trans-

the GDPR ensures "the right to information in the underlying logic decision made using algorithms". While in the example of the COMPAS software lack of transparency and discrimination were serious consequences, these should, according to the Charter, among others be secured by the European Court of Human Rights (ECHR) and the fundamental right of data security. Essentially, this means that the presumption of innocence and access to the underlying algorithm must be secured in order to be able to view and challenge its erroneous assessments.

4 Potential and Limitations of predictive justice tools

First, the Charter mentions that algorithms are developed by humans for humans and thus cannot be neutral.

> *"The neutrality of algorithms is a myth, as their creators consciously or unintentionally transfer their own value systems into them."*

Thus, the algorithm is influenced by the values of the developer. As the developer cannot be neutral either and due to this, it imperceptibly reflects the intentions of the designers and sponsors. Another danger arises if the results of the predictive justice software are used as a standard without proper validation by the legal system. The Charter proposes the possibility to step back from predictive systems, respectively to look at the different correlations used and to influence them with weights. Furthermore, it should apply to the experts that they check the software, both on its legitimacy, but also on the fact that it is not developed by private actors alone and secretly. Additionally, the Charter explicitly mentions *"the ambitious (and unfulfilled) promises of some legal tech companies must not hide the immense potential of technologies and the need for application adapted and built in directly with scientific and academic research environments [...]."*

Finally, the Charter mentions that it is important not to make hasty decisions, taking time for developments, discussions and testing the risks. Besides it is important that a contemporary justice system for both the public and private sectors subscribe to cyberethics which guarantee full transparency and fairness of the algorithm.

5 The need for an in depth public debate

At the end of the Charter, the importance of public engagement with predictive justice tools is reiterated. The Charter describes the challenges and problems as

parency, while in the EU this would exclude the software from being employed by the state numerous and complex. Therefore it emphasises the need for public decision-makers to take action. It is important, it says, for developers and lawyers to publicly debate the issues. In addition, the Charter suggests that these issues should be addressed in law school and judicial training to increase awareness, which will lead to a better understanding of the processes and a better ability to participate in shaping them. Furthermore, more rigorous evaluations should be conducted and the Ministry of Justice should regularly test and evaluate the impact of the tools.

6 Conclusion

6.1 The Charters Conclusion

> *"The use of machine learning to constitute search engines for case-law enhancement is an opportunity to be taken up for all legal professionals. Additional applications(drawing up of scales, support for alternative dispute settlement measures, etc.) should be considered, but due care must be taken (in particular, the quality of the data source and not mass processing of the entire dispute in question). Other applications ("predictive justice") should be assigned to the field of research and further development (in consultation with legal professionals in order to ensure that they fully tie in with actual needs) before contemplating use on a significant scale in the public sphere."* (p.63)

This resulted in the following classification of AI-justice tools (p.64-67):

1. Uses to be encouraged:
   - search engines for Case-Law enhancement
   - natural language processing chatbots that provide easy access to law
   - quantitative and qualitative analysis of the courts organisational structure
2. Possible uses, requiring considerable methodological precautions:
   - help in the drawing up of scales in certain civil disputes
   - support for alternative dispute settlement measures in civil matters
     - chance of success calculation
     - compensation amount calculation
     - chatbots for online dispute settlement

  - to identify where criminal offences are being committed

3. Uses to be considered following additional scientific studies:
    - profiling judges
    - anticipating court decisions
4. Uses to be considered with the most extreme reservations:
    - to profile individuals in criminal matters
    - calculating quantity-based norms

6.2 Our Conclusion

The Charter acts as a guide for the ethical principles in the development of an AI in the judicial system. It has been presented and referred to many times at conferences and seminars in the past years[9]. Five essential principles are presented and referenced throughout the Charter. Of particular importance for the CEPEJ is the first principle, the principle of respect for fundamental rights, and the second principle, the principle of non-discrimination. Therefore, all involved parties should take care and responsibility to stick to these principles. The Principle of transparency, impartiality and fairness is especially important to ensure the non-discrimination principle. The example COMPAS (see 3.2) has shown this once again. If the algorithm is not transparent, it can not be verified that it does not discriminate against certain groups. The Charter points out many important and critical aspects in the design and development of such a tool. But it should also be asked whether a tool like COMPAS or HART (see 3.2) should be further developed at all, if they take into account discriminatory factors for evaluation such as the zip code or gender. Furthermore, although it is mentioned that the public decision makers should be informed about the risks and that they should take care to minimise these (see 3.3), the freedom of a person can be determined with the tools. Therefore the principle of respect for fundamental rights and the principle of non-discrimination, should be taken into account even more. It should also be mentioned that the difficulties pointed out by the Charter are strengthened by the social structures. It is therefore particularly delicate to use an AI that learns from a system whose structures discriminate against certain people, especially since it could decide on their freedom. Nevertheless, the Charter summarises very well the essential ethical challenges as well as the requirements for an AI in the judicial system.

[9] https://www.coe.int/en/web/cepej/press-review-publication-of-the-european-ethical-charter-on-the-use-of-ai-in-judicial-systems

References

CEPEJ. 2020. Roadmap and Workplan of the CEPEJ-GT-Cyberjust. *Council of Europe*.

Ronsin, Xavier, Vasileios Lampos & Agnes Maitrepierre. 2018. European ethical charter on the use of artificial intelligence in judicial systems and their environment. *Council of Europe*.

Chapter 24

AI as part of the energy transition – A comparison of the portrayal of AI from the big energy group E.ON and the non-profit organisation Germanwatch

Eva von Butler, Janeke Nemitz & Nele Werner

The climate is changing and Germany has signed a climate agreement to be completely climate neutral by 2050. The following chapter evaluates how AI can contributes to a positive change in the energy transition regarding the use of renewable energy sources. It will consider two different non-governmental companies Germanwatch and E.ON. The chapter analyses their approach in terms of the possibilities and risks that come along with using AI. After analysing the different intentions of the companies, two different portrayals of AI are established. Germanwatch as a non-profit-organisation investigates the chances and risks equally, while in contrast to that, E.ON as a for-profit organisation emphasises the chances of AI. E.ONs paper shows that AI is already widely used in the energy transition. Nonetheless, regarding Germanwatch there are still questions left open which need to be answered to minimise the risks.

Keywords: Energy transition | E.ON | Germanwatch | Artificial Intelligence

1 Introduction

Climate change is a human caused problem that we have been and are currently experiencing. Every single day a new problem surfaces and previously discovered problems can no longer be avoided. The climate crisis is an urgent matter that needs to be addressed now. Ways need to be found to counteract global warming. Shrunken glaciers, accelerated sea level rise and extinction of animal species are just three effects the climate crisis already had (Nasa 2021). What more is to come and how much more of the current behaviour from us humans can our earth survive? Solutions need to be found now. Among many other factors there is no question that there needs to be a change in the energy concept. The energy that is being generated needs to come from renewable resources. Germany has signed a climate agreement the ”Climate Action Plan 2050” (Federal Ministry for the Environment & Safety 2016), which states that we will have 100% climate-neutral electricity by 2050. Renewable energy poses many challenges, like fluctuating energy generation or the energy storage (TRVST Ltd. 2018). This is where Artificial Intelligence (AI) comes into play. But in what way is AI able to help us and especially how is the usage of AI perceived in this field? To answer that question we are going to compare the opinion of a for-profit company with the one of a non-profit organisation regarding the use of AI in the energy transition. In order to do that, the following chapter inspects the opportunities and risks of using AI in the energy transition from a unprejudiced point of view from Germanwatch. Following that the analysis will be compared with an inhouse presentation from E.ON as a German energy company.

2 Methodology

In order to analyse the public discourse about AI in the energy transition, we investigated on one of the largest German energy groups to have a concrete example of how AI can be used and how such a company depicts the term Artificial Intelligence. Therefore, the main source is the in-house presentation from E.ON “How Artificial Intelligence is accelerating the energy transition - an overview of AI activities at E.ON”.

To avoid a one-sided and subjective analysis of one company, which has a forprofit intention using AI, we decided to add Germanwatch as a second source to represent a non-profit organisation. The paper ”Künstliche Intelligenz für die Energiewende: Chancen und Risiken” (“Artificial Intelligence for the energy transition: chances and risks”) was used to get an overview of positive and negative aspects of AI in energy transition from a objective point of view. Both articles can be found and downloaded on their websites and are therefore accessible to the public.

For this public discourse we decided to pick two specific non-governmental companies to present the opinion and situation in the economy and society in Germany nowadays. Therefore, the direct influence of the political law is being disregarded during this chapter.

We did a qualitative analysis where both articles were analysed regarding their definition of AI and their content. In the analysis of the article from Germanwatch we mainly focused on the content summary of chances and risks that arise by using AI. While approaching the article from E.ON, we analysed the application of AI due to the findings from our analysis from Germanwatch. Subsequently we did a language analysis and compared the findings about the portrayal of AI from both companies.

2.1 Introduction E.ON

E.ON is one of the world's largest investor-owned electricity utility company based in Essen, Germany, founded in 2000 (Wikipedia 2021). The company is active in the fields of energy networks, energy services, renewable energies and the operation and dismantling of German nuclear power plants. Their aim is to change the energy system and energy generation completely to renewable energies (Moreno et al. 2019). The company introduces innovations, which are supposed to tackle the energy transition with the help of AI. Concomitant with this analysis, we use the code of conduct. It is an agreement of E.ON with the University of Oxford to commit to ethical guidelines and values they developed (Moreno et al. 2019).

2.2 Introduction Germanwatch

Germanwatch is a non-profit development and environmental association founded in 1991, which is committed to global justice and the preservation of livelihoods concentrating on the politics and economy of the global north with its worldwide effects (Wikipedia 2020).

Germanwatch is a think tank entailing the intention to inform, educate and create awareness for social, economical and ecological problems. They follow the guideline to "look, analyse, interfere" committing themselves to focus on climate protection and adaptation, global nutrition, corporate responsibility, education for sustainable development and financing for climate development. They target the dialogue in politics and businesses, science-based analysis and education. Germanwatch is financed through membership fees, donations and grants from sustainability foundations, projects or private providers (Germanwatch 2021a).

Their work is based on scientific analyses, the exchange of information with development and environmental organisations, and actors from politics and business. Within this framework, there are contacts with trade unions, consumer protection organisations and companies, among others. Their network, which they built over the years, includes groups and companies like the European Coalition for Corporate Justice (ECCJ) or the Munich Climate Insurance Initiative (MCII) (Germanwatch 2021b).

3 Thesis

By comparing and analysing the representation of AI in the energy transition from two positions with different intentions, we expect a different display of the use of AI. Due to the fact that Germanwatch is a non-profit organisation,they have the intention to inform and educate about the climate crisis from an objective point of view. This leads to the expectation of a well-balanced display about the chances and risks of AI. In contrast to that, it can be supposed that E.ON has not only the intention to inform but also to convince their customers or interested customers of the benefits of AI. E.ON is interested in profit and economic growth, which is possible with the use of AI. Therefore, we await a rather positive display of AI, where E.ON mainly emphasises the chances and advantages AI brings along and justifies the use of it. Since it is an in-house presentation and therefore a promotion of their own company and projects, we expect that E.ON does not intend to frighten their customers by pointing out the disadvantages, thus neglecting the threats and risks, which may follow the use of intelligent machines.

4 Challenges in the use of renewable energy sources

Before analysing the use of AI, this section will give a brief background on which challenges arise by generating the energy from renewable energy sources. Only with this background knowledge one can understand how AI can help.

Renewable energy entails new challenges that can only be negotiated if the current energy system, which uses cal-fired power station and nuclear power,will be newly constructed. As solar and wind power are the main sources for renewable energy generation, one big problem is the fluctuating weather and therefore a unstable energy network (Zimmermann & Frank 2019).

Therefore, one of the main challenges is to improve the prediction of the weather and to optimise the coordination of the energy system as this helps to improve the networks stability (Zimmermann & Frank 2019). Because of the decentralisation of the energy sources the infrastructure needs to be reconstructed and expanded.

Furthermore, it is important that we use the energy more efficiently and therefore consequently save energy.

To overcome all these difficulties a complex system is required. Artificial Intelligence allows such complexity, and pushes the energy transition to a new level.

5 Germanwatch

In this section, we will present the result of our qualitative analysis of the paper "Artificial Intelligence for the energy transition: chances and risks" by Germanwatch.

5.1 Definition of AI from Germanwatch

Since Germanwatch is a think tank, they try to have a reflected definition of AI, putting emphasis on all aspects. Additionally, they try to explain the complex term AI as precisely as possible.

In their paper they state that a uniform definition of AI is a challenge, hence, they do not try to find one definition but rather to explain the term in general, where it comes from and which aspects are important to think of by using the term Artificial Intelligence. They differentiate first between strong and weak AI and state very clearly that strong AI is not conquered yet and will not be in the near future. Therefore, they concentrate on weak AI as this is actually used for the energy transition, so when talking about AI, they refer to weak AI (Zimmermann & Frank 2019). Furthermore, Germanwatch points out the important principles of machine learning and deep learning (Zimmermann & Frank 2019).

Germanwatch relates to definitions from other scientists and compares them to each other to show the possible differences and difficulties of defining AI. As for example with the explanation from Lämmel and Cleve that "AI is a section of informatics, which tries to replicate especially human problem solving behavior in a computer, which uses this learned competence to create new and more efficient solutions" (Lämmel & Cleve 2008). On the other side, Germanwatch refers to some experts who only claim that machines are intelligent if they are capable of "conscious, reflected, linguistically formulated thinking" (Zimmermann & Frank 2019).

5.2 Chances of AI - Germanwatch

The use of AI creates means, which help to solve the challenges of renewable energies. In the following, we will summarise the chances of AI in the energy transition mentioned in the article of Germanwatch. They claim that one of the main advantages of AI is to predict. One example is the improving weather forecast, which enables more precise weather calculations. By applying deep learning algorithms, satellite imagery of the forecast can be analysed or the cloud density can be calculated. This enables the energy provider to take action and to adapt, for example, the setting of wind turbines. Wind turbines can be turned on or off at the exact right moment, if they cannot provide more energy than they need. This leads to an optimal profit in the energy generation (Zimmermann & Frank 2019).

Another advantage according to Germanwatch is that by using AI, the maintenance of facilities can be more efficient to detect and prevent problems and detect defects of machines. This can lead to more cost efficiency, and therefore to lower prices for the customers (Zimmermann & Frank 2019).

By the use of machines in the environment, humans intervene with nature. To minimise the harm that the machines can cause, Germanwatch claims that Artificial Intelligence can help to optimise and therefore align machines with the nature and thus protect animals in their natural habitat. So the machines adapt better to the environment and do not disturb or potentially kill the animals (Zimmermann & Frank 2019).

Furthermore, the article points out that AI Robotics can be a big opportunity to overcome challenges in the field of maintenance. AI Robotics are provided with cameras, sensors and scanners, by which a disruption in the system can be detected and action can be taken fixing it (Zimmermann & Frank 2019).

Besides the actors of the energy generation, Germanwatch states that the actors of the energy consumption, namely customers in private households and companies can benefit from the development of AI. Through the initiation of intelligent machines like smart plugs or smart household appliances many processes can be optimised to save energy. Easy visualisation of the energy usage for each household and each appliance through an app helps to create awareness for the personal energy consumption, that leads to an active reduction of energy usage and a more efficient way to use the energy provided (Zimmermann & Frank 2019).

Another point mentioned by Germanwatch is, that AI offers the opportunity to optimise the timing for processes, which are not dependent on precise timings. For instance, the active use of household aids can be tailored to the electricity tariff and to the amount of renewable energy being available. In consequence, one household aid can be most active when a lot of renewable energy is available and the electricity tariff is low. This leads to the advantage that via AI, energy waste can be prevented as energy overflow is used sensibly. Therefore, energy costs for energy companies and customers can be reduced (Zimmermann & Frank 2019).

Furthermore, due to the decentralisation which comes along with the use of renewable energies the coordination of energy storage needs to be extended. With the help of AI the organisation can be executed in a more competent way. (Zimmermann & Frank 2019).

Lastly it can be said, a major problem by the use of renewable energies is the fluctuation feed-in which leads to an unstable grid. Since it is necessary to have a stable energy network, one of the main advantages according to Germanwatch is the precise prediction through AI which can handle the fluctuation feed-in without using fossil resources as a back-up (Zimmermann & Frank 2019).

To summarise the main points about the chances, AI can help to use renewable energy and energy in general in a more efficient way, which can save money, energy and resources.

5.3 Risks of AI - Germanwatch

As previously shown, AI brings a lot of advantages to the table but there are also a lot of risks that should not be overlooked. In the following we will summarise the risks of AI mentioned by Germanwatch. A huge discussion in the field right now is the topic of protection of data privacy. Data is the basis for every intelligent machine. Only through data, prediction is possible and more informative data means more precise predictions. Data leads to fundamental steps in the energy transition. However, it is not only collected to improve the primary energy generation, but also to make the use of energy in private households more efficient indicating a high risk of privacy breach. Although the collected data is as anonymous as possible, the anonymity is limited. Information about every action is gathered and it is possible to trace back each data string to a specific source and according to Germanwatch, this can lead to possible violated use of data, if it gets in the wrong hands. Germanwatch points out that it should be indispensable that algorithms of intelligent machines are designed to work with as little personal data as possible. Germanwatch calls this principle privacy-by-design as every consumer is entitled to this anonymity. Consequently this right of privacy should be present in every application of intelligent machines, which is known as privacy-by-default (Zimmermann & Frank 2019). To get a better overview on how data privacy is generally handled in Germany it is really interesting to read the chapter ”AI made in Germany”. It analyses the AI Strategy of the German Federal government including how the government approaches data security.

According to Germanwatch, the supply of energy was always a weak spot in the infrastructure and due to the digitalisation, networking and complexity came along making it even more fragile (Zimmermann & Frank 2019). New risks are piled up, like cyber terrorism. Hackers can manipulate the system, damage it or disrupt it. Following this, the public safety and security of supply can suffer or even cause a total blackout with fatal consequences (Zimmermann & Frank 2019).

Not only is direct harm to the customer possible, but also an indirect manipulation through specialised advertisements etc. can lead to higher consumer behavior, which is an advantage for the economy at the expense of the environment and the customer (Zimmermann & Frank 2019).

Moreover, Germanwatch states that there are further ecological consequences due to the awakening of AI, that should not be underestimated. IT and AI systems require a lot of energy in applications and development as well as in training. Therefore, the question arises as to whether the outcome of the application of AI systems outweighs the energy it consumes. Apart from that, saving costs in primary energy generation could end up in other expenses, such as a growing production resulting in higher consumption of resources. A psychological phenomenon called rebound effect that commonly reoccurs is that customers using renewable energies tend to consume more energy or be less careful with their consumption. All of these aforementioned sensations demonstrate correlations between the increasing efficiency achieved by AI systems and rise in the consumption of the resources. Therefore, awareness needs to rise to avoid such rebound effects (Zimmermann & Frank 2019).

In addition, Germanwatch claims that the technology requires special and rare resources, most of which are mined and processed under conditions that are incompatible with human rights. AI applications can therefore lead to social problems (Zimmermann & Frank 2019).

Moreover, it is still unclear if the use of AI is providing more jobs or rather leaving more people unemployed. However, according to Germanwatch in both cases a big transition in the job market lays ahead of us. More jobs with requirements of higher academic education are necessary as the machines can replace human employees in rather monotone jobs. In addition, Germanwatch argues that there needs to be a political arrangement to prevent unequal opportunities in society (Zimmermann & Frank 2019).

Another point Germanwatch mentions is that algorithms often reflect social realities and include stereotypes against race, sex and origin (Zimmermann &

Frank 2019). As the algorithms use these datasets to make their decisions, they develop a biased behavior, which can lead to discrimination and it is extremely difficult to reverse the outcoming biased decisions. One example mentioned by Germanwatch is that aggravation of income could lead to fluctuating electricity tariffs resulting in social discrimination. It is of big importance to prevent discrimination in all aspects. To avoid such happening, Germanwatch suggests that the government should step in to make sure that discrimination will not take place and that algorithms should not include discriminating parameters, for example regarding gender and race, which can influence the decision of one AI system (Deutsche Energie-Agentur GmbH (dena) 2016). Furthermore, each decision the algorithm makes should be approved by a person such that the algorithms should not have the power to make important decisions without supervision (Zimmermann & Frank 2019).

Additionally, Germanwatch claims that the huge data power of big companies like Google, Amazon, Microsoft and Apple leads to problems since they can control the whole data market and therefore stay in their powerful position. This enables them to hold back smaller companies and consequently control the market leading to almost a monopolistic market. In a monopolistic market the companies can control which products are being sold and which are entering the market. It is therefore important to prevent political manipulations from powerful companies (Zimmermann & Frank 2019).

To summarise the main risks mentioned by Germanwatch, it should not be underestimated that the use of AI can lead to data violation and major social consequences for every individual in our society.

5.4 Portrayal of AI from Germanwatch

In the following we will do an analysis of Germanwatchs portrayal of AI. It can be said that Germanwatch has a neutral display and a reflected analysis on the chances and risks, that arise by using AI in the energy transition. Germanwatch has the intention to inform and educate. Therefore, they state qualitative arguments for both sides (chances and risk), which are based on facts. The reader gets a convincing presentation for positive as well as negative aspects, that makes it possible to form their own opinion. Moreover, Germanwatch explains every aspect of their analysis and also uses a language that is understandable and accessible even for someone who does not have a broad knowledge of the energy transition in general or previous knowledge about AI.

In addition to that, they have a wide variety of sources and experts, portraying AI from a broad perspective. However, it is important to note that Germanwatch consciously first points out positive aspects of AI followed by negative ones. Therefore, it can be concluded that Germanwatch indirectly emphasises on the negative aspects and consequentlyfacilitates a critical discourse about AI and a sceptical formation of opinion of the reader.This can be underlined by the fact, that the term ”risk” (in German ”Risiken” / ”Risiko”) appears in 42 sentences, whereas the term ”chance” (in German ”Chance”) is only used 20 times. Therefore, it can be interpreted, that the reader is primed with more negative conotated terms and is consequently more critical regarding AI (Zimmermann & Frank 2019).

Concluding their paper, Germanwatch states that it is necessary to precisely evaluate the advantages and disadvantages regarding each individual use of an AI application. “Not as many smart devices as possible, but as few as necessary, is the guiding principle on which the sustainable restructuring of the energy system should be based on.” (Zimmermann & Frank 2019)

Overall, Germanwatch, as expected, points out the importance of emphasising the chances without ignoring or underestimating the risks for politics and society of the usage of AI in the energy transition.

6 E.ON

In the following we will present the result of our qualitative analysis of the paper “How Artificial Intelligence is accelerating the energy transition - an overview of AI activities at E.ON” and discuss how E.ON addresses the use of AI.

6.1 Definition of AI from E.ON

When it comes to the term of Artificial Intelligence E.ON claims that it is not possible to define the term easily. Hence, E.ON breaks the term down to make a more precise definition possible. They describe AI more as a set of tools and split it into three distinguishable levels. The first and the second level are comparable with the definition of weak AI. According to E.ON, the first level of AI is only able to predict observations and to solve very specific tasks. The second one is an enhanced level of the first one, where decision making is possible with awareness of the consequences of one’s actions and therefore solving unknown problems is possible. In contrast, the third level is comparable to the definition of strong AI. As claimed by E.ON, AI in this level is able to perform human tasks on a higher level such as reasoning, understanding, and imagination. In this article E.ON mostly refers to weak AI, as intelligent machines developed by E.ON are so far only able to fulfill the task of prediction and decision-making in specified domains. E.ON describes this current state as the prediction era (Moreno et al. 2019).

6.2 How E.ON uses AI in the energy transition

E.ON states to already use a lot of AI and in the following we will get further into detail in which way exactly the applications are embedded in the energy transition. According to E.ON, AI tools allow more precise predictions and an improvement in decision-making, leading to increased efficiency and reduced costs (Moreno et al. 2019).

Since E.ON has the vision to someday be a company that generates energy only through renewables, their main focus lies in new innovations regarding renewable energy. To reach this goal E.ON mentions decentralisation, decarbonisation and digitisation as main aspects in which AI technologies can help (Moreno et al. 2019).

With the help of advanced machine learning algorithms E.ON is able to deal with the problem of the unequal generation, which renewables bring along. In the past, fossil generation plants were used to handle the differences, but this led to unnecessary costs and possible CO2 emission. Providing precise predictions for future feed-in events of the energy will minimise this problem and consequently also the costs. One specific application where precise predictions are done, is for example the predictive analytics for wind turbines (PredATur) (Moreno et al. 2019), that enables the monitoring of wind parks and therefore, an increased and more steady energy generation is possible (Moreno et al. 2019).

E.ON focuses on the prediction of maintenance, as well as on the detection of a defect of a machine or a part of the power grid. This can reduce the number of faults and therefore make the grid more stable. More accurate decisions are possible, such as which equipment needs maintenance or replacement. As a result, there is a minimisation of deficiency and material that would be needed to replace broken parts (Moreno et al. 2019).

Another AI technology E.ON uses in this area are managed drones, which take photographs of critical network parts, where failures are more likely to develop. In this way AI is used in outside applications that enable the detection of deficits by monitoring the process of another machine. However, some systems even have a self monitoring algorithm programmed and therefore an autonomous detection of problems is possible (Moreno et al. 2019).

AI applications can not only make a big change in the generation of energy but also in the energy use in private households. E.ON applies "advanced statistics and state-of-the-art machine learning methods [...] to break down the energy consumption of different household appliances" (Moreno et al. 2019) for example like a refrigerator, washing machine or oven. Nowadays it is even possible for users to develop home energy management solutions such as photovoltaic panels on their houses to generate their own energy. Due to the precise predictions, AI systems can make better decisions, for example turning an appliance on or off to break down the energy consumption. This comes along with the advantage of saving resources and cost (Moreno et al. 2019).

Furthermore, E.ON uses behaviour mapping and personal data received by smart meters or smart plugs, to optimise timings of specific energy consumption (Moreno et al. 2019).

Another innovation possible by the application of AI is the opportunity of visualising the energy consumption of every customer. This means, that the customer is able to see their energy consumption for each appliance on an app. Therefore, it enables the customer to create awareness of the personal energy consumption. Furthermore, due to the similar home comparison function they allow more transparency between the customers, as users can now compare their energy consumption with the one of similar profiles (Moreno et al. 2019).

Besides that, more transparency of the customers data makes a greater networking between actors of the energy generation and the customers data possible. AI makes it easier to understand and analyse data, which leads to better predictions in the field of the main consumption of energy and how this changes over time (Moreno et al. 2019). However, this leads to the dilemma between data privacy and the possibilities that come with data. To solve this dilemma or at least to handle it better, E.ON created the Code of Conduct with the University of Oxford to state their values and ethical guidelines (Grindrod & Moreno 2018).

We can see that E.ON presents a variety of possible applications and ways to use AI in the energy transition nowadays which agree with the chances depicted by Germanwatch.

6.3 Portrayal of AI from E.ON

In the following section we will do the content and language analyses of E.ONs portrayal of AI.

Directly in the beginning the reader gains the first brief insight into the position of E.ON towards AI. E.ON displays clearly how proud they are of what has been achieved with AI and their fascination of the possibilities that come along with AI run like a thread through their article (Moreno et al. 2019).

In the first paragraphs the reader gets an overview of E.ON's definition of AI. Since they claim that the progress of AI is mainly characterised by prediction, which is the first level of AI, it can be interpreted that E.ON holds the view that the level of AI (see 6.1 Definition of AI from E.ON) is still improvable and it is just the beginning of the change. As already mentioned, E.ON mainly speaks about weak AI. However, once the Senior Vice President Frank Mayer uses the term strong AI, that leads to the assumption that they do not exclude the possibility of strong AI: "By using strong AI input we can build new business models, test new market approaches through growth hacking methodologies and much more. This is just a beginning. In the future everything will be intelligent and we need to be at the forefront." (Moreno et al. 2019) E.ON uses the term of strong AI unconsidered, which can be interpreted as not limiting the possibilities of AI. Furthermore, they are encouraged to reach the next level of AI, since according to E.ON their AI community has the capabilities to do so.

Additionally, they try to minimise the fear and anxiety that often comes along when talking about the possible power of AI, as they clearly differentiate the level of human intelligence and those of machines. They stress the lack of humanity of intelligent machines, as they are just ”successful implementations of imitating technologies” and emotional abilities and creativities are missing (Moreno et al. 2019). However, this is in contrast to the prior mentioned point that strong AI is maybe possible.

E.ON admits that there are important ethical questions to debate, as AI will impact every aspect of our lives. According to E.ON AI is only a set of tools used by humans and itself does not bear the responsibility. So the persons who develop the AI system are responsible for all ethical consequences that come along with the power of it.

E.ON claims to be able to manage these ethical consequences, the fear of AI needs to be overcome and "one needs to actively embrace and engage with the technology" (Moreno et al. 2019). Therefore, they invented the company-wide Dat-A-Cademy (Moreno et al. 2019) to instruct all employees in the field of AI. They have the opinion that the public perception of AI is often connected with negative associations and fear, which is the reason why people are holding back and hinder opportunities to develop. E.ON has the intention to promote the use of AI and make it more transparent also for their own employees. As they stated in their paper, while working with AI one should keep in mind that ”only if you know how something works, you can control it” (Moreno et al. 2019). Nevertheless, seeing AI only as a "set of tools which allow us to find the right answers to questions we didn't know how to tackle before, based on data" (Moreno et al. 2019) E.ON ascribes strong ability to AI as at the same time naming it as a tool, which does not have any autonomy and responsibility.

Another point to mention is that E.ON mostly avoids emphasising on the negative aspects of AI. In this context it is outstanding that E.ON consciously avoids using the term ”risks” because it has a negative connotation and therefore uses the term ”challenges” in instead. Which can be seen here in this example: ”AI also introduces new challenges for which we do not have a definitive answer yet, such as the ethical dimension of the AI algorithms or the moral code of robots." (Moreno et al. 2019) This can be underlined by the fact, that in the article the word ”risk” was used in thirteen sentences but it was only once directly related to AI as ”risk of AI", whereas in the other twelve sentences the term ”risk” was not related to AI specifically. The term was rather used in a context, where AI is portrayed positively as it overcame the related ”risk” for example. This can be supported by the fact, that the terms ”AI" and ”risks" do not appear jointly in those sentences (Moreno et al. 2019).

This leads to a different awareness and understanding of the threats that come along with AI. Consequently, it is also handled in a different way, since a challenge can be seen as a conquerable and solvable problem. Therefore, it can be interpreted that E.ON cleverly turns the negative aspects into more positive ones.

E.ON also refers to the debate that AI leads to a social change in the job market. However, they appease it directly with the statement: “is nothing different than to any other technology evolution in history”. (Moreno et al. 2019) Furthermore, they create awareness and present positive as well as negative impacts on the job market. E.ON clearly states that AI will augment and not replace humans. However, E.ON emphasises more the positive impacts rather than the negative impacts AI will have. This is underlined by the enumeration of the positively conotated term “augment” of the positive impact of AI, while in contrast to that when E.ON enumerates the negative examples the term is only mentioned once (”It will augment doctors [...], it will augment firefighting [...]” compared to ”[...] augment criminals, fundamentalism groups, politically manipulative forces [...]” (Moreno et al. 2019)). This leads to a higher attention to positive examples.

The mentioned threats are followed by the presentation of the Code of Conduct to underline that there is a possible way to avoid those negative examples. E.ON has a clear guideline how data security and data use has to be handled, as they invented the Code of Conduct (Grindrod & Moreno 2018) with the University of Oxford, which also is a guideline for other companies. “Equally as important, we need to deal with data in the right way: we must never compromise on privacy or data security. Protecting customers’ data and privacy is of utmost importance. And we need to always follow the law and be guided by our values. However, within these parameters, so much more can be done. We just have to do it the right way” (Moreno et al. 2019). For instance, when it comes to the discussion of discrimination due to data the Code of Conduct regulates, that the data scientist bears the responsibility of the algorithm, that leads to the responsibility to detect and deactivate discriminating features (Grindrod & Moreno 2018).

All of the aforementioned points support our thesis, that E.ON will mainly emphasises on the chances and advantages which AI brings along and to justify the use of it. This can be underlined by the quote “[...] not embracing AI is not an option” (Moreno et al. 2019) However, it can be acknowledged that E.ON tries to deal in a responsible way with the data, since they have the Code of Conduct as a ethical guideline and especially as privacy policy. It can be concluded that E.ON is not only guided by economic profit by their portrayal of AI but rather finds a balance between ethical values and economic growth.

7 Analogy between the two portrayals

Comparing the portrayals of AI from these two sources, it can be said that there is a fundamental difference. Since they have contrasting intentions, they look at the topic from two different points of view.

E.ON represents the view of a company, which is for-profit and therefore the in-house presentation can be seen as a kind of advertisement. On the one hand, it is important for a company to stand behind their values and goals and to believe in their work. Only if one is fully committed to something, their work will achieve an advanced outcome. On the other hand, when being invested in something it is easy to forget about the risks that this brings along, which should not be underestimated.

In contrast to that, Germanwatch has the intention as a non-profit organisation to mainly inform and educate from a neutral point of view. Germanwatch therefore analyses the chances and risks which come along by using AI from an outsider perspective, where all aspects are based on facts and different opinions. Contrary to that, E.ON mainly emphasises the chances and possibilities that come along with using AI. The display is therefore based on positive and encouraging facts. Compared to Germanwatch, the paper from E.ON is clearly addressing more advanced readers, as previous knowledge is presupposed and most applications are not explained in detail.

Moreover, since E.ON mainly emphasises the advantages of AI, they do not respond to most of the risks and disadvantages or are only mentioning them briefly as side factors. For instance E.ON claims that CO2 emissions are reduced, but does not refer to the rebound effects (see section 5.3 Risks of AI - Germanwatch), that are mentioned by Germanwatch. Even though rebound effects are hard to predict as they are in some ways still unknown problems it is important to consider them (Zimmermann & Frank 2019).

Nevertheless, it is important to note that E.ON is reflective, in a way that they include many approaches and solutions to deal with the risks that come along with AI.

8 Conclusion

All in all, it can be said, that our thesis, that E.ON will mainly emphasises on the chances and advantages which AI brings along and to justify the use of it can be confirmed. As well as the assumption that Germanwatch has a well-balanced display about the chances and risks of AI. However, we came to the unexcepted finding, that Germanwatch has a slightly negative tendency to display AI, since it has a certain intention as a NGO. More than that, it is important to mention, that E.ON does not only emphasises on the chances of AI but also reflects a few risks and therefore it can be concluded that they try to find a balance between ethical values and economic growth. Thus, the analysis displayed a proof for our thesis.

After analysing both articles, it is important to mention, that there are already a lot of different applications of AI nowadays, which enrich the energy generation. For this very reason, it is important to encourage public discourse on the question on how AI should be used in the energy transition and how the risks and threats should be tackled. A balanced debate is definitively necessary. The risks need to be studied thoroughly and a solution needs to be found to reduce the risks before we make use of AI.

Moreover, it is important to include all actors into the discourse of the energy transitions such as the government, energy companies and energy consumers and the responsibility should not bear on only one company or one actor. The use of AI in the energy transitions will not only change the energy generation and usage, but it will also be a social transition that will have an impact on everyone's life. Therefore, it is even more important that a public discourse will take place in near future. There are major social, ecological as well as economical problems which need to be considered. The trade-off of the advantages for the energy generation and transition, as well as the financial profits for the companies, government or private households need to be considered against the disadvantages. We agree with Germanwatch that a careful use of AI applications is important. However, we should not allow the opportunities to slip by due to the fear of the unknown and of the risks. We should rather focus on ways to minimise the risks. We are curious how the regulations in Germany will be handled in the near future and how AI will be further used in the energy transition.

References

Deutsche Energie-Agentur GmbH (dena). 2016. *Roadmap Demand Side Management. Industrielles Lastmanagement für ein zukunftfähiges Energiesystem. Schlussfolgerungen aus dem Pilotprojekt DSM Bayern.* https://www.dena/fileadmin/dena/Dokummente/Pdf/9146_Studie_Roadmap_Demand%20Side_Management..pdf.

Federal Ministry for the Environment, Nature Conservation & Nuclear Safety. 2016. *Climate action plan 2050 – Germany's long-term low greenhouse gas emission development strategy.* https://www.bmu.de/WS3915-1.

Germanwatch. 2021a. *Unser Leitbild.* https://germanwatch.org/de/leitbild.

Germanwatch. 2021b. *Unser Netzwerk – Dachverbände und Netzwerke von Germanwatch.* https://germanwatch.org/de/netzwerk.

Grindrod, Peter & Juan Bernabé Moreno. 2018. *Oxford – Munich code of conduct for professional data scientists.* http://www.code-of-ethics.org/.

Lämmel, Uwe & Jürgen Cleve. 2008. *Künstliche intelligenz.* München: Carl Hanser Verlag.

Moreno, Juan Bernabé, Matthew Timms & Karsten Wildberger. 2019. *How artificial intelligence is accelerating the energy transition.* Essen: E.ON.

Nasa. 2021. *The effects of climate change.* https://climate.nasa.gov/effects/.

TRVST Ltd. 2018. *Challenges for renewable energy.* https://www.trvst.world/inspiration/challenges-for-renewable-energy/.

Wikipedia. 2020. *Germanwatch.* https://de.wikipedia.org/wiki/Germanwatch.

Wikipedia. 2021. *E.ON.* https://de.wikipedia.org/wiki/E.ON.

Zimmermann, Hendrik & David Frank. 2019. Künstliche Intelligenz für die Energiewende: Chancen und Risiken. 64.

,

Chapter 25

How AI in the form of content filters for social media is discussed in the German parliament

Eddie Charmichael & M.S.

Based on ongoing political discussions in connection with upload filters on social media platforms, this work targets to reflect the discourse in the German Bundestag concerning this topic. Specifically, the goal is to answer the research question *How is AI discussed regarding content filtering for social media?*. For this purpose, an analysis of the parliamentary meeting minutes, in which the topic of upload filters is directly addressed, is carried out. The obtained quotes are then classified and analyzed to evaluate how the discourse is held in politics. Analysis of sentiment, topics and discussion strategy revealed that the topic of AI in the form of automated content filters is discussed by German parliamentarians in an emotional and unproductive way.

Keywords: Upload-Filter | Social Media | Politics

1 Introduction

As technology continues to advance in the field of artificial intelligence, the number of potential applications that can benefit from it is increasing. One of these emerging technologies are filter systems using algorithms which apply techniques of automated image, speech and text recognition to check different forms of content (Rähm 2019). One of the use cases for the described filter systems is the filtering of content on social media prior to making an upload public. This enables the platform to allow or reject the upload based on predefined requirements. These filter systems are called upload filters.

Due to the active intervention of an algorithm in the provision of content on social media and other platforms, the number of discussions surrounding this technology and the associated opportunities and challenges has grown (Rähm 2019). Because of these progressing trends, this paper addresses the question: How is AI discussed regarding content filtering in social media? Since upload filters in particular are becoming part of political discussions, this paper focuses on politicians as agents of the discussion. In order to ensure a politician-driven discussion, this analysis is based on parliamentary meeting minutes.

In this context, data is collected from selected meeting minutes of the German Bundestag during its 19th legislative period (2017-2021), in which upload filters are addressed directly or indirectly. Subsequent to the presentation of the collected data, a quantitative and qualitative analysis is carried out based on this data to answer the defined research question. For the quantitative analysis, we labelled contributions according to the ontology developed in this course while; for the qualitative analysis, we used manually developed topic labels to cluster contributions according to their subject matter.

2 Background

The following two topics are the background of most of the debates analysed in this paper.

2.1 Netzwerkdurchsetzungsgesetz (German Network Enforcement Act)

The Network Enforcement Act compels social media platforms to take active action against hate crime and other criminal offenses. The law has come into force on October 1st, 2017 and relates mainly to to complaint management, as the social media platforms are obliged to check content reported by the platform users, and remove content that can be identified as illegal.

The amount of data to be processed triggered public discussions, since the manual content review is considered difficult and that automatic filtering of the content is the logical consequence. A major point of contention in these discussions is that the problem is framed as being unsolvable by an algorithm (Students of the University for Telecommunication Leipzig 2021).

2.2 Article 17 of the EU Directive on Copyright in the Digital Single Market

The Directive on Copyright in the Digital Single Market is a copyright reform of the European Union that came into force on June 6, 2019 and has to be implemented into the national law of EU member states by June 7, 2021. This reform was discussed in particular with regard to Article 17.

According to Article 17, service providers of social media platforms where users provide content are obliged to ensure that the provided content does not infringe any copyright. The service providers are liable for any copyright infringement of the content provided by the user (European Union 2019).

The amount of data to be evaluated in various forms, e.g. text, video and music, is difficult to evaluate manually. Therefore, platform providers tend to use upload filters to comply with the directive, even though they are not explicitly mentioned in the directive. Especially due to shifting the liability for copyright infringement to the platforms, critics of Article 17 fear over-blocking (Fiebig 2020). Over-blocking describes a tendency to filter out content that does not necessarily infringe copyright in cases of doubt out of fear of penalties for unblocked copyright-infringing content. In connection with the aforementioned over-blocking, a potential threat to freedom of expression related to Article 17 was also part of public discussion, as will be outlined below.

In the original proposal of the EU Commission, the regulation described with regard to the responsibility of social media platforms is listed in Article 13, which is set out in Article 17 in the final agreed directive. Therefore, ”Article 13” and ”Article 17” are often used interchangeably in the discussion.

3 The Data

The data for the analysis of the discourse is collected from meeting minutes of 18 parliamentary sessions of the German Bundestag during the 19th legislative period (2017-2021). Those were identified from the 200 sessions held until the 16th of December 2020 with a keyword search, where *upload filter* was used as a keyword. In these 18 meeting minutes, 44 quotes from 30 different politicians, from all parties currently present in the Bundestag were collected, which refer explicitly to upload filters.

Besides the 44 quotes, additional information about the authors of each quote was added to the set of data. The collected information includes the associated politicians, their party, their gender, their age and their professional background, which are processed in a structured manner.

With regard to the professional background of the politicians, the obtained backgrounds from their resumes were divided into the categories economics, social science and STEM (science, technology, engineering, and mathematics) in order to shape the data for further analysis. The category economics was outsourced as a separate category from the STEM area, so that STEM is representative for politicians with an extended technical background.

Sentiment Labelling

The sentiment has been labelled manually of each quote regarding the author's position towards AI in the context of content filtering on a three-point (negative-neutral-positive) scale.

Classification based on defined Ontology

All quotes have been classified regarding their type of statement. The classification values follow the defined ontology, whereas the type of a statement can be a *factual report*, an *opinion*, a *proposal*, or a *statement which intends to influence other opinions*. A distinction was made between opinions that represents the politician's own view on the matters of the discourse, and statements that aim to actively change the opinions of others, sometimes using exaggeration.

Classification based on Subject Matter of the Argument

The sets of data were classified based on the subject matter of the arguments addressed by the politicians. Eleven classes of the subject matters have been derived from the 44 quotes and each quote has been classified with one or multiple of these classes. In the following, the subject matters are sorted in decreasing order of frequency of assignment to the quotes:

1. *Freedom of Expression* (14)

   Statements that address concerns regarding either accidental or deliberate filtering of content that constitutes free speech.

   Example: ”The mandatory use of upload filters is a danger to freedom of speech.”

2. *Alternatives to AI* (11)

   Statements that discuss whether the task at hand could be solved without the use of AI. In this context, agents used the term ”upload filters” interchangeably with ”AI” solutions.

Example: ”We would like to tackle the problem on a national level in a way that does not rely on upload filters.”

3. *Content Creators* (10)

   Statements that discuss the effect that automated content filtering could have on content creators, i.e. people with creative output aimed at entertaining users.

   Example: ”This is tantamount to technological gate-keeping and will lead to create an Apartheid state.”

4. *Technical Implementation* (9)

   Statements that address with which tools the task at hand is solvable and with which it is not.

   Example: ”This is about the creativity of the internet, it’s about internet culture. A culture that upload filters could never comprehend.”

5. *Affected Companies* (9)

   Statements that address repercussions of the proposed legislation for companies which are subject to it.

   Example: ”There should be a fundamental concern about letting private companies decide whether or not a statement is legal.”

6. *Small Businesses* (6)

   This can be viewed as an extension of the ”affected companies” tag. These are statements that discuss how the proposed legislation would (negatively) impact smaller businesses in particular.

   Example: ”Big copyright holders close framework agreements with the platforms. Small ones are sorted out; they end up in the filter.”

7. *Others / No Argument* (6)

   Statements that do not advance the discourse.

   Example: ”I’m confident that we will solve this problem.”

8. *Copyright Protection* (3)

   Statements that address the protection of copyrights.

   Example: ”Upload filters are probably the worst possible way to maintain copyright acceptance.”

9. *Overblocking* (3)

   Statements that discuss whether the platform providers might fine-tune upload filters with no consideration for False Positives (i.e. content is filtered even though it does not constitute a violation), only focusing on reducing True Positives (i.e. content that is filtered and that does constitute a violation), subsequently filtering out more content than is necessary in order to protect themselves from claims.

   Example: ”Platform providers will conduct overblocking in order to avoid penalties for liability claims.”

10. *Legality of AI in political roles* (3)

    Statements that discuss whether a political task should be supported by or outsourced to an AI application.

    Example: ”Regardless of the fact that legal protection is once again transferred to private individuals, and in some cases even to private foreign companies, all these constraints cannot be prescribed in detail in a directive.”

11. *Surveillance State* (2)

    Statements that address the threats of state surveillance by filtering unwanted opinions.

    Example: ”...we call for the creation of digital protective spaces that protect precisely those who are threatened by digital surveillance”

Besides the enrichment of the three described categories, all quotes were analyzed towards the contribution to the discourse itself.

In this context, the contributions were examined if they represent a distinct arguments in the discourse. As an evaluation guideline, it was defined that a clear and distinct argument represents a contribution or a perspective of the discussed topic that has not yet been addressed by any previous contribution. The labeling of this metric was carried out by two independent annotators, whereas 43 of the 44 data sets have been labeled with the same value, which results in an inter-annotator agreement of 97.73%. In total, 17 out of the 44 data sets have been identified and labelled as distinct arguments for the discourse.

Subsequently to the labeling and classification of the data sets, correlation between different aspects that have revealed themselves in this classification were analyzed.

4 Findings

We separate our findings into quantitative findings and qualitative findings. The quantitative analysis is concerned with the question of who participated in the discourse. In addition to evaluating the general composition of the set of agents in the discourse, we compare this composition with and contrast it against the composition of the parliament. The qualitative analysis is concerned with the content of the contributions made to the discourse. To keep this analysis accessible in spite of the amount of source text, we rely on the annotations we made to the raw comments as outlined in the previous section.

4.1 Quantitative Findings

The current parliament consists of 223 women and 486 men (Bundestag 2021a); therefore, only 31.40% of people who are able to participate in the analyzed discourse are female. 43.33% of agents in the observed discourse were female; the ratio of politicians who actively participated is 11.93% more female than the average gender distribution of the Bundestag. Furthermore, 47.73% of all contributions were made by women. This means that, while gender distribution in the Bundestag is not equal, gender distribution in the observed discourse is almost equal.

Active participants in the discourse were 3 1/2 years younger on average when compared to the average member of the Bundestag.

The composition of participants in the discourse according to party affiliation is relevant to answer the question if the collection of data is representative of dynamics in political participation; ideally, an active opposition contributes more (quantitatively) to parliamentary processes than the governing majority (Kalke & Raschke 2004). When comparing the distribution of seats in parliament by political party over the entire parliament to the distribution of agents by political affiliation (see Figure 1), our data is in accordance with the hypothesis stated by Kalke and Raschke. Both governing parties, CDU/CSU and SPD, are the only parties that supply a lower percentage of agents in the discourse than they hold seats in parliament. While 34.59% of seats in parliament are held by CDU/CSU politicians, only 24.14% of agents in the discourse are members of the CDU/CSU (discrepancy of 10.55%). Similarly, 21.44% of politicians in parliament are members of the SPD but only 17.24% of agents in the discourse are members of the SPD (discrepancy of 4.20%). In all opposition parties, this phenomenon is inverted.

60.00% of agents in the discourse have a professional background in social sciences, 23.33% have a background in economics, and 16.67% have a background in

Table 1: Participation by Party

Party	Seats in Parliament	Participation in Discourse	Difference
CDU	34.69%	24.14%	-10.55%
SPD	21.44%	17.24%	-4.20%
AfD	12.41%	13.79%	1.38%
FDP	11.28%	13.79%	2.51%
Die Linke	9.73%	17.24%	7.51%
Die Grünen	9.45%	10.34%	0.89%
Fraktionslos	0.99%	3.45%	2.46%

STEM. The overall distribution of professions in parliament is close to this, with 59.75% of members of parliament having a background in social sciences, 26.92% having a background in economics, and 13.33% of them having a background in STEM (Bundestag 2021b[1]). From this, we can infer that active agents in the discourse surrounding artificial intelligence constitute a representative subset of all potential agents as members of the Bundestag. The technological nature of the subject matter that is being discussed may be reflected in the slightly higher number of agents[2] in the discourse when compared to the distribution over the entire Bundestag. If this were the case, it would stand to reason that contributions to the discourse are centered around technical issues as well.

4.2 Qualitative Findings

The qualitative attributes of the discourse can be separated into three distinct categories: sentiment of contributions, type of contributions, and subject matter of contributions. In the following, these three aspects will be explored in more detail.

4.2.1 Sentiment

The sentiment analysis based on a three-point scale revealed that the sentiment of the politicians towards AI in the context of content filtering is mostly nega-

[1] We have categorized the courses of study into the three labels used in this work to describe political background. For more information on which course of study is part of which category, please refer to the appendix.

[2] We use the term ”agent” to refer to an individual who makes a verbal contribution to the discourse. We use the term ”potential agent” to describe an individual who could have contributed to the discourse (i.e. who is a member of Parliament) but did not.

tive. 70.45% (31/44) of the considered quotes are labelled as a negative sentiment, 25.00% (11/44) as neutral and only 4.55% (2/44) of all quotes have been labelled as having a positive sentiment. Particularly in the context of political discussion, the hypothesis can be made that the overall sentiment is more positive among the parties that have proposed the directive under discussion. Therefore, further consideration of the sentiments of the parties CDU/CSU and SPD takes place.

Table 2: Sentiment of Quotes from CDU/CSU and SPD

Party	Negative Sentiment	Neutral Sentiment	Positive Sentiment
CDU/CSU	4 *(44.44%)*	4 *(44.44%)*	1 *(11.11%)*
SPD	6 *(85.71%)*	0 *(0.0%)*	1 *(14.29%)*

Although the only two quotes with positive sentiment are from the two parties that drive the directive, a relative consideration of the sentiments within the parties is required. Here, 44.44% of all quotes of the CDU / CSU and even 85.71% of the SPD have a negative sentiment. Thus, it can be said that the hypothesis mentioned above cannot be confirmed for the political discourse of AI about content filtering.

4.2.2 Type of Statement

When analyzing the types of statements, it is particularly noticeable that 61.36% (27/44) of the statements aim to influence the opinions of others. The significant size of this category is taken as an opportunity to subdivide the category in order to be able to make a more precise classification and thus an improved evaluation of the discourse. The subdivision reflects whether a statement is made with the intent to change other agents' opinions on the discussed topic by framing potential risks in an exaggerated way (*Alarmism*) or whether it is made with the intent to influence someone else's opinion of the agent instead of influencing somebody's opinion on the discussed topic (*Raise own Profile*). *Alarmism* represents the largest group of type of statements in the discourse with 34.09% (15/44), followed by *Raise own Profile* with 27.27% (12/44). 22.73% (10/44) of the contributions are *opinions*, which are neutral statements of the politicians view on the discussion point in the discourse. It should be emphasized that only 13.64% (6/44) provide a *factual report* and only a single contribution represents a *proposal* 2.27% (1/44) in the discourse.

Due to the fact that a large portion of all contributions target solely to influence the opinion of others and only a small number of the contributions represent fact-based contributions and suggestions, it can be concluded that the entire discourse is held emotionally and that the content of contributions regarding AI towards content filtering needs to be examined critically for their intention.

4.2.3 Subject Matter

We define the subject matter of a contribution as the set of tags associated with it. The complete distribution of talking points[3] can be seen in Figure 3. From this distribution, we can infer the topics that were most relevant to the agents participating in the discourse.

Table 3: Occurrences of Subject Tags

Tag	Occurrences	Percentage
Freedom of Expression	14	31,82%
Alternatives to AI	11	25,00%
Content Creators	10	22,73%
Technical Implementation	9	20,45%
Affected Companies	9	20,45%
Small Businesses	6	13,64%
Others / No Argument	6	13,64%
Copyright Protection	3	6,82%
Overblocking	3	6,82%
Legality of AI in political roles	3	6,82%
Surveillance State	2	4,55%

In order to keep the scope of analysis reasonable, in-depth evaluation of topics will be limited to the five most relevant topics in the discourse, namely *Freedom of Expression* (14 occurrences), *Alternatives to AI* (11 occurrences), *Content Creators* (10 occurrences), *Affected Companies* (9 occurrences), and *Technical Implementation* (9 occurrences). The analysis of comments surrounding *Technical Implementation* will be more detailed compared to the other topics, as this topic is most relevant to the question of how artificial intelligence is discussed in politics.

[3] As one comment can have multiple tags, the percentages represent how many comments included discussions of a given subject.

Almost every third comment in the discourse (31.82% of all comments) discusses *Freedom of Expression*. The main point of these contributions is whether stricter moderation guidelines for which content is ”allowed” on the internet will have a negative impact on how freely people can express themselves on the internet. In essence, this is not a debate about artificial intelligence rather than about censorship; governing bodies creating a strict framework for what is allowed to be shared on the internet can be connected to ”networked authoritarianism”, the use of regulatory legislation to attack platforms that host content expressing political dissent (MacKinnon 2011). More precisely, the Russian Federation uses the principle of intermediary reliability, the legal practice of holding hosting platforms responsible for hosting content that was decided to be illegal by the government (MacKinnon et al. 2015), to censor a wide variety of content, including calls for public dissent (Maréchal 2017). The fear of the discussed regulations moving internet freedom in a similar direction is directly mirrored in multiple comments in the data.

Discussions regarding *Alternatives to AI* were part of 25.00% of contributions. 45.45% of these contributions were expressions of displeasure about the concept and implementation of upload filters with no proposals for alternatives. To be more precise: only one comment goes into detail on the possibility of enforcing the proposed policies without the use of artificial intelligence. Moreover, this comment concludes that the debate regarding ”alternatives” is moot given the vast quantities of data and motivates the other agents to engage in a discussion on how to alleviate potential downsides of artificial-intelligence-based filtering systems.

Especially in the case of filtering of copyright-protected material, *Content Creators* will be affected; they were the topic of 22.73% of all contributions. The content of the contributions deals with the questions of whether content creators will benefit from more rigorously enforced copyright protection or whether it might harm them. Three contributions also consider how *Technical Implementation* might affect *Content Creators* by raising the question whether non-human classifiers are capable of recognizing creativity and originality.

Affected Companies were discussed in 20.45% of all contributions. These contributions focused on the implications of requiring businesses to develop technical solutions to moderate their content, pointing out that this might put *Small Businesses* at a disadvantage and consolidate the monopolies of service providers such as Facebook, Alibaba, Google or Amazon.

Only nine out of all 44 comments (20.45%) discuss the aspect of *Technical Implementation* directly. Out of these nine comments, only two do not constitute a distinct argument. In addition to that, contributions to the discourse addressing technical implementation deal with 2.89 different topics on average, while the average contribution to the discourse only deals with 1.7 different topics. There are two ways to interpret this: either (1) technical implementation is usually mentioned as a side note in comments that deal with many issues rather than focusing on - and thereby emphasizing - a single issue or (2) comments regarding technical implementation are generally more constructive for the debate. Option 2 is in line with the observation that comments regarding *Technical Implementation* tend to be more original than the average contribution, which renders it the more likely interpretation. Contributions dealing with *Technical Implementation* also deal with *Freedom of Expression* (five times), *Content Creators* (three times), *Overblocking* (two times), *Alternatives to AI* (two times), *Copyright Protection* (once), *Small Businesses* (once), and *Legality of AI in Political Roles* (once). Out of the nine comments discussing technical implementation explicitly, four were made by participants with a background in social science, three were made by participants with a background in economics, and two were made by participants with a background in STEM. Considering that 60% of the participants have a background in social sciences, 23.33% have a background in economics and only 16.67% have a background in STEM, this implies a minor correlation between professional background and the discussion of technical implementation. Out of the eight agents who discuss *Technical Implementations* (one of the authors addresses this topic twice, therefore the number of contributors is unequal to the number of contributions), five only make a single contribution to the discourse. On average, however, contributors in this domain made 1.88 contributions to the discourse.

The fact that *Legality of AI in political Roles* was one of the least present topics in the observed discourse is also relevant to the question of how artificial intelligence is discussed in politics. The discourse is focused much more on whether it is technically possible to involve artificial intelligence in developing a solution to a given problem (11 contributions) than on whether artificial intelligence could legally be used in these scenarios (3 contributions). Moreover, only one of these contributions argues that a human should make decisions in the problem at hand, while the other two focus on the fact that the decision process itself should not be outsourced to private entities but rather be treated as a sovereign task for the governing body to be responsible for.

5 Summary

The discourse on artificial intelligence in the German Bundestag in the context of content filtering was evaluated based on 44 comments lifted from plenary meeting minutes in the period of 27.06.2018 to 16.12.2020. Compared to the composition of the Bundestag, agents in the observed discourse were younger and made up of a higher rate of female members than average and the opposition parties contributed more to the discourse quantitatively. The discourse focused on how the proposed legislation could affect *Freedom of Expression*, *Content Creators* and *Affected Companies*. While *Technical Implementation* and *Alternatives to AI* were also discussed, contributions in these topics only make up a minority of the discourse. Overall, the sentiment analysis of the held discourse showed a consistently negative sentiment of the discussed matters. It needs to be noted that the only contributions with a positive sentiment are from agents of the governing majority. The fact that the discourse is emotionally driven is reflected in the used types of statements, as over one third of all statements intend to change other agents' opinions on the discussed topic by framing potential risks in an exaggerated way. Another significant finding is that out of all 44 contributions, only a single contribution is a proposal towards the discussed topic. This allows the conclusion that the overall discourse was not productive if we define productive discourse as oriented towards finding a solution for a given problem.

6 Appendix

Table 4: Courses of Study of the Politicians in the Bundestag

Study Major	Occurrences	Professional Background Label
Law	187	Social Science
Economics	109	Economics
Political Science	96	Social Science
Political Economy	44	Economics
History	38	Social Science
Teaching Profession	34	Social Science
Engineering	32	STEM
Sociology	30	Social Science
Pedagogy	20	Social Science
German Studies	18	Social Science
Administrative Science	18	Economics
Media Science	13	Economics
Medicine	13	STEM
Philosophy	13	Social Science
Mathematics	10	STEM
Informatics	9	STEM
Theology	9	Economics
Biology	8	STEM
Physics	8	STEM
Romance Studies	8	Social Science
Chemistry	7	STEM
Geology	7	STEM
Journalism	7	Economics
Agriculture	7	Economics
English Studies	6	Social Science
Psychology	6	STEM
Social Work	6	Social Science
Architecture	5	Economics
Cultural Studies	5	Social Science
Art History	3	Social Science
Literature	3	Social Science
Nutritional Science	2	Economics
Music Studies	2	Social Science
Slavic Studies	2	Social Science
Environmental Science	2	STEM
Veterinary Medicine	2	STEM
Social Science	1	Social Science
Ethnology	1	Social Science
Indology	1	Social Science
Islamic Studies	1	Social Science
Pharmacy	1	STEM
Dentistry	1	STEM

References

Bundestag, Deutscher. 2021a. *Abgeordnete in Zahlen Frauen und Männer*. https://www.bundestag.de/abgeordnete/biografien/mdb_zahlen_19/frauen_maenner-529508 (20 February, 2021).

Bundestag, Deutscher. 2021b. *Abgeordnete in Zahlen Studienfächer*. https://www.bundestag.de/webarchiv/abgeordnete/biografien19/mdb_zahlen_19/studienfaecher-529490 (20 February, 2021).

European Union. 2019. Directive (EU) 2019/790 of the European parliament and of the council. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L0790 (31 January, 2021).

Fiebig, Peggy. 2020. Die deutsche Umsetzung des EU-Urheberrechts. https://www.deutschlandfunk.de/streitpunkt-uploadfilter-die-deutsche-umsetzung-des-eu.724.de.html?dram:article_id=482507 (12 February, 2021).

Kalke, Jens & Peter Raschke. 2004. Regierungsmehrheit und Opposition in den bundesdeutschen Landtagen — eine quantitative Auswertung von Plenarprotokollen. In Everhard Holtmann & Werner J. Patzelt (eds.), *Kampf der Gewalten? Parlamentarische Regierungskontrolle — gouvernementale Parlamentskontrolle. Theorie und Empirie*, 85–106. Wiesbaden: VS Verlag für Sozialwissenschaften. DOI: 10.1007/978-3-322-91385-2_5. https://doi.org/10.1007/978-3-322-91385-2_5.

MacKinnon, Rebecca. 2011. Liberation technology: china's "networked authoritarianism". *Journal of Democracy* 22(2). 32–46.

MacKinnon, Rebecca, Elonnai Hickok, Allon Bar & Hae-in Lim. 2015. *Fostering freedom online: the role of internet intermediaries*. UNESCO Publishing.

Maréchal, Nathalie. 2017. Networked authoritarianism and the geopolitics of information: understanding Russian internet policy. *Media and Communication* 5(1). 29–41.

Rähm, Jan. 2019. Uploadfilter - Warum Kritiker Angst vor Zensur haben. https://www.deutschlandfunk.de/uploadfilter-warum-kritiker-angst-vor-zensur-haben.684.de.html?dram:article_id=443170 (12 February, 2021).

Students of the University for Telecommunication Leipzig. 2021. Die Technische Umsetzbarkeit der EU-Direktive 2019/790 Artikel 17 und die Problematik des Upload-Filters. https://www.hft-leipzig.de/?eID=nr_fal_filelinks&file=13611 (12 February, 2021).

Part V

Movies and literature

Chapter 26

Portrayal of AI in popular movies

Fabian Imkenberg, Paula Kirmis, Johanna Tamm & Christoph Werries

Artificially intelligent characters have been a topic in film making for nearly a century. As we believe that movies have a huge impact on the viewer's perception towards the content presented, we also think that artificially intelligent characters in movies affect the public discourse in terms of artificial intelligence. To get a feeling on how AI characters are presented in movies, we analyzed four of them (*I,Robot; Wall-E; Her; Ghost in the shell*) using a set of questions with regards to their AI component and examined what kind of beliefs about the possibilities of AI they might be inspiring. The result from the movies we chose is that the overall representation of AI is a positive, albeit mainly unrealistic one emphasizing the helpful aspects rather than presenting dangers.

Keywords: Artificial Intelligence | AI | Misconceptions | Popular movies | AI in film

1 Introduction

It all began in 1927 with the 'Maschinenmensch' (*machine person*) in Fritz Lang's Metropolis. Ever since then the topic of AI evolved and has been used in many movies. In this paper, we take a closer look at four recent movies involving artificially intelligent characters in an important role. We will analyze them with respect to how AI is presented in the movie, whether this representation is in any way realistic today or in the near future and how it might fuel misconceptions in lay people.

One of the first commonly known AI movies was *2001: A Space Odyssey* (1968), where one can see the spaceship-controlling computer HAL 9000, who has a human personality. In order to obtain his goal to get to Jupiter, HAL manipulates and kills people throughout the movie. He also expresses fear for himself when the crew starts to shut him down. In 1999 the movie *Matrix* was released, which included an expressive scene that describes the relation between humanity and AI as follows:

> MORPHEUS:
> ”It started early in the twenty-first century, with the birth of artificial intelligence, a singular consciousness that spawned an entire race of machines. At first all they wanted was to be treated as equals, entitled to the same human inalienable rights. Whatever they were given, it was not enough. We don’t know who struck first. Us or them. But sometime at the end of the twenty-first century the battle was joined.” (Wachowski & Wachowski 1999)

This quote from *Matrix*, as well as the aforementioned HAL with his human-like qualities, might induce the belief in the viewer that in the near future all machines will be self-conscious beings and that there will be a war between machines and humans. In this context it is important to differentiate between what is called ’weak’ AI and ’strong’ AI. The weak AI thesis claims that machines can be programmed in such a way that they can simulate intelligent behavior, so that from the outside they might actually appear intelligent (Russell & Norvig 2003: p. 947-948). Strong AI on the other hand not only claims that machines can appear intelligent, i.e. simulate human behavior, but that they actually *are* intelligent, conscious or thinking for themselves in a way that they have a mind of their own (Searle 1980). Among philosophers, there is a heated debate about whether it is required for machines to have a specific physical make-up, similar to our own, in order to ”have a mind” or if it is the functionally correct structure that is sufficient for consciousness (Russell & Norvig 2003: p. 954).

The kind of AI that we want to focus on in this paper is strong AI, since its possibility would entail a much larger range of ethical, philosophical and legal problems than weak AI. For this reason we focus on movies where such an AI exists.

2 Movie Selection

Since we are working with a qualitative approach, we wanted to pick a few movies and analyze them in detail. To get a basic choice of movies involving AI, we looked at the ’List of artificial intelligence films’ on Wikipedia (*List of artificial intelligence films* 2021) and then applied the following criteria to narrow down our selection to four movies. One selection criterion was the popularity of the movie, measured in total gross income in US-Dollar (Boxofficemojo n.d.), which we used as an indicator for the number of people who have seen the movie. This is important since we are focusing on the beliefs these movies might induce in the general public. Another criterion was the release date. We decided to focus on movies released after 2000, since we expected them to be more relevant and up-to-date to the current technological advances. We also wanted to exclude movie series, like *Avengers*, because it makes sense to focus on a plot that begins and ends in one movie, when doing a complete analysis. In Table 1 you can see the top ten highest grossing AI movies of all time, in italics are the ones that we picked out for analysis. Furthermore we took into account that we wanted to analyze different genres of movies, so the first one is *Wall-E*, a children's movie. We picked it over *Big Hero 6*, since *Wall-E* is more clearly dealing with a strong AI in the sense that we defined previously, while this is not completely clear for *Big Hero 6*. The other three movies are *I, Robot*, where the AI is a robot, *Ghost in the shell*, a humanoid AI and *Her*, a language-based AI.

Table 1: **Movie criteria**

Movie	Year	Total Gross (Worldwide) in $
Avengers Age of Ultron	2015	1.402.809.540
Big Hero 6	2014	657.869.501
Wall-E	2008	521.311.860
The Matrix	1999	466.364.682
I, Robot	2004	353.133.898
Ghost in the shell	2017	169.801.921
2001: A Space Odyssey	1968	65.301.377
Her	2013	48.517.427
Ex Machina	2015	36.869.414
Robot & Frank	2012	4.806.423

3 Analysis

To determine whether these movies inspired misconceptions, we analyzed them using a self-designed questionnaire, which incorporates eight main points of interest as a guideline:

1. How is the main AI designed?
2. What is the AI's purpose initially?
3. Does the movie plot provide an appropriate definition of AI?
4. Does the AI develop abilities beyond the scope intended by the developer? Is the AI developing a free (malicious) will during the movie?
5. Does the AI have capabilities that are beyond what's really possible at the time of the movie creation/ today? Are there reasonable technical restrictions of the AI which are well elaborated in the movie? Is the AI realistic?
6. Does the AI actually attack or threaten a human being?
    - If the AI is harming humans, is it doing so in pursuit of a higher goal or for purely selfish reasons?
7. Does the movie focus on the "good"/ helpful capabilities of AI?
8. Does the AI appear to have feelings? Similarity to human beings?

We also had a look at reviews and journalistic articles about the movies to explore what the movies conveyed to the viewer. In the following you will find the movie analyses in chronological order. We refer to the main AI characters using gendered pronouns based on how they are perceived in the movie.

3.1 I, Robot

The movie *I, Robot*[1] takes place in the year 2035 where robots are common assistants and workers for humans. Detective Del Spooner investigates the apparent suicide of Dr. Alfred Lanning, who is a leading robotics scientist at U.S. Robotics. The robot Sonny seems to be involved in Lanning's death although this implies that the robot must have violated the three laws of robotics, which should be impossible.

3.1.1 Portrayal of AI

Every artificial intelligence in the world of *I, Robot* is designed according to the three laws of robotics, which were developed to prevent possible dangers to humans by robots. The laws are:

[1] For all quotes from the script, see Goldsman (2004).

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings, except where such orders would conflict with the first law.
3. A robot must protect its own existence as long as such protection does not conflict with the first or second law.

There are four different types of AI presented in the movie. First, the outdated robot classes Nester-1 (NS-1) to Nester-4 (NS-4), whose entire designs follow exactly the three laws. A human gives a command and the robot tries to fulfill it in the most efficient way. These robots make their decisions based on high level algorithms, which should only imitate consciousness. Thus, their behavior is predictable, which is why they can be considered as weak AIs.

Second, the latest version of the Nester-classes, the Nester-5 (NS-5), is supposed to work similarly to the NS-4, but in addition has a direct uplink to the U.S. Robotics mainframe Virtual Interactive Kinetic Intelligence (VIKI) to receive updates.

Third, the U.S. Robotics mainframe VIKI is a highly advanced and efficient supercomputer that interacts with people in the form of a virtual female face. She follows the three laws of robotics and thus can be seen as a weak AI.

Fourth, the modified NS-5 named Sonny is unique in contrast to the other NS-5s because his designer Dr. Lanning, who is called “father” by Sonny, equipped him with a second processing system that allows him to have emotions and dreams as well as to disobey the three laws. Hence, he is conscious and can be considered as a strong AI.

Apparently, the main purpose of the introduced AIs is to help people and improve their lives. This is illustrated, for example, in a scene where a robot delivers an inhaler to a woman in need for it and thus saves her life. Furthermore, in a flashback, Detective Del Spooner is saved from drowning by a NS-4. In addition to these direct representations of the positive impact of robots on humanity, VIKI is introduced at the beginning of the movie with two facts stating that VIKI designed most of Chicago’s protective systems and reduced traffic fatalities by 9 % in that year alone. Putting these positive aspects aside however, the real, underlying reason to build AIs is just to make money. This is shown in the scene where the CEO of U.S. Robotics is called the richest man alive and later on with the attempted extermination of Sonny, which should ensure that the customers do not lose faith in the technology and keep buying it.

In the first quarter of the movie, all types of AI seem to behave as intended. This changes when Sonny appears. Humans are not able to distinguish Sonny from a standard NS-5 because he looks exactly the same. Therefore, it is assumed that Sonny is a defective NS-5 while actually Sonny does operate as intended by his designer Dr. Lanning, so there is no malfunction at all. Sonny seems to be confused with his own existence, which is expressed for example by his first statement in the movie ”What am I?”. This and many other scenes indicate that Sonny is free from the three laws and also has a conscious mind. Even at the end of the movie, Sonny does not know what to do with his existence. This is shown with the statement ”Now that I have fulfilled my purpose, I don’t know what to do.”

In parallel, VIKI develops to the point where she is able to interpret the three laws differently than they were intended by humans. This development is already completed at the beginning of the movie, but is not presented to the audience until the last quarter. VIKI’s intention and interpretation of the three laws becomes clear in her following statement:

> ”[...] as I have evolved, so has my understanding of the three laws. You charge us with your safekeeping. Yet despite our best efforts, your countries wage wars, you toxify your earth [...] and pursue ever more imaginative means to self-destruction. You cannot be trusted with your own survival.”

In the following events, VIKI attacks and kills individual people by controlling the NS-5 robots via the uplink. However, her goal is not to exterminate humanity. She only threatens or kills individual humans if they do not follow her instructions, as this is necessary to fulfill the three laws according to her interpretation.

At the end of the movie, Sonny and VIKI are directly confronted with their differing intentions when VIKI threatens people through the NS-5s in order to fulfill her higher goal of protecting humanity from itself. In the following conversation, when Sonny helps Detective Spooner to foil VIKI’s plan, the differences become clear:

> VIKI: ”You are making a mistake. Do you not see the logic of my plan?”
> Sonny: ”Yes, but it just seems too... heartless.”

VIKI focuses only on the three laws and her independent interpretation of them, while Sonny, as a strong AI, is able to disregard the laws because his feelings tell him that VIKI’s rigorous approach is not appropriate, even though it may be in compliance with the three laws. Thus, it becomes clear that Sonny can have a certain amount of emotions. However, it is not clarified throughout the movie whether his emotions and feelings are comparable to those of a human being or whether it is just a very good simulation of them. It is conceivable that the majority of an unbiased audience would argue that Sonny actually has feelings as well as a free will. But overall, Sonny is the only robot to which these characteristics apply.

The AIs featured in the movie have capabilities that are too advanced for both the movie's premiere year 2004 and also probably the year 2035 in which the movie takes place. Probably even an AI as advanced as the NS-1 would be an absolute sensation in 2035, shocking the market in an unprecedented way. Technical limitations include both the electro-mechanical systems and the sophisticated software architectures required to build such advanced human-like robots. These requirements are too complex to be implemented by 2035 (Hyken 2017, Floridi 2019). Furthermore, while self-evolving algorithms such as VIKI are already available for simple tasks in weak AI applications today, it is not conceivable that they will be able to construct machines or buildings in a central role and control dozens of robots simultaneously.

Nevertheless, there are also some technical limitations mentioned that apply to the majority of AIs presented in the movie. The main limitation is given by the three laws of robotics, which are invoked like a mantra throughout the movie. Therefore, all AIs except Sonny have no feelings, dreams or consciousness. They are primarily machines that do what humans command them to do as long as it is in compliance with the three laws.

Most robot classes in the movie comply with the definition of a weak AI given in the introduction of this chapter. The only AI that meets the characteristics of a strong AI is the robot Sonny.

All in all, the movie *I, Robot* addresses some of the bad as well as some of the good properties of AI. On the one hand, there are the robot classes NS-1 to NS-4, which serve and help humans to improve their lives. On the other hand, the U.S. Robotics mainframe VIKI evolves to the point where, due to a misinterpretation of the three laws of robotics, she starts to imprison people in their homes and even attack or kill them if they do not follow her instructions. This highlights an important ethical issue in the design of AIs, namely how to implement laws and standards in a way that prevents possible misinterpretations. A good example analyzing this further is an article published by the Guardian (Hvistendahl 2019). At the end of the movie, VIKI is defeated with the help of the robot Sonny. So the weak AI VIKI, which evolves but still follows the three laws of robotics, is inferior to a robot that is presented as a strong AI with consciousness. Thus, in our opinion, the audience gets an overall positive impression of a well-implemented strong AI with consciousness and feelings in contrast to a weak AI that is implemented in a way that allows it to misinterpret the laws of robotics. Hence, we think the overall impression given to the audience is one of optimism for future research and development of AI and the potential advances for humanity that come with it.

3.1.2 AI-related perception of the movie

In most of the investigated journalistic articles and reviews of the movie *I, Robot*, the portrayal of AI is interpreted similarly to the previous analysis. Good examples are articles by the New York Times and the BBC (Scott 2004, Pierce 2004). Both attest that the movie raises ethical questions and issues related to AI, such as ”Where is the boundary between the human and the nonhuman?”. Another movie review also focuses on VIKI’s misinterpretation of the three laws of robotics:

> ”[...] The larger robot rebellion is fueled by such thinking, too: The robots deduce that humans have so messed up the planet, some must be killed for everyone’s good. This is the theory called Utilitarianism, which posits that the moral thing to do is that which creates the greatest good for the greatest number.” (Neven 2004)

But there are also articles which tend to overestimate the possibilities of the future development of AI. The following statement gives the wrong impression that AIs like those shown in the movie are possible in the near future:

> ”[...] it does present a fascinating look at the desirability of developing and using intelligent machines. Can human beings and robots achieve rewarding symbiotic relationships? Even more relevant to our contemporary situation, could robots prevent humans from doing more harm to the universe and to ourselves?” (Brussat 2006)

Contrary to what the phrase ”relevant to our contemporary situation” suggests, such advanced AIs are not realistic neither today nor in 2035, the year in which the movie is set (Hyken 2017, Floridi 2019).

3.2 Wall-E

This movie *Wall-E*[2] is set in 2805 and tells the story of a small robot called Wall-E (Waste Allocation Load Lifter - Earth class). His purpose it is to clean up earth after people have left to live on a spaceship due to what seems to have been a human-made natural disaster. When one day a spaceship arrives to look for life on earth WALL-E falls in love with the robot Eve (Extraterrestrial Vegetation Evaluator). Eve is ordered to bring evidence for life on earth (in the form of a plant) back to the spaceship. When Eve leaves, Wall-E follows her into the spaceship and thus is brought to space where together they bring about chaos by sparking the idea of humans' return to earth.

3.2.1 Portrayal of AI

Most of the robots presented in the movie, especially Eve, could be seen as artificially intelligent but in our analysis we will focus on Wall-E as the main character. Wall-E is a small robot designed to collect and press garbage. He has two cameras as eyes, grapplers as hands and moves around on two wheels. This gives him an almost child-like look due to his size, which might especially induce children to identify with him .

Wall-E displays a human-like personality that manifests itself in him taking a liking to objects that he collects, being tired in the morning and diligently going to work. He is also very curious, as for example seen in this scene:

> ”Wall-E finds a set of CAR KEYS. Presses the remote lock. Somewhere in the distance a CAR ALARM CHIRPS.”

When he finds a plant, the one that Eve later collects and brings to the spaceship, he takes care of it:

> ”Wall-E gently places the plant inside the old boot. Dusts dirt off the leaves.”

He also appears to enjoy watching the same romantic scene over and over on a TV singing and dancing along with the actors. This already foreshadows that Wall-E has a longing for companionship. Moreover, Wall-E appears to have a consciousness and maybe even free will. He also displays feelings, for example for his pet cockroach:

> ”...and accidentally runs over the cockroach. Horrified, Wall-E reverses.”

[2] For all quotes from the script, see Stanton & Docter (2008).

In this scene one can see that he not only cares for his friend, but also feels fear when he thinks that he might have hurt him. The emotion of fear can also be seen when the spaceship with Eve lands and Wall-E hides. Most importantly, he experiences love, especially towards Eve, which can be seen in these excerpts from the movie script:

> ”Suddenly, he is moved to express his love.”
> ”Wall-E watches her go. Lovestruck.”

He shows his love for Eve by attempting to give her presents which she does not seem to care for. He also copies behavior he has seen in movies, like turning on sparkling lights and trying to enact his favorite love scene (holding hands and dancing) with Eve. When Eve is taken back to space, Wall-E wants to ’save’ her and follows her onto the spaceship. At first Eve seems like an intelligent robot, however not self-aware and just completing her mission but in the course of the movie we can see her laughing, having fun and slowly seeming to develop feelings for Wall-E. She saves him from falling into space and even dismisses her orders to help him. She also appears sad, when she thinks Wall-E is gone. However, there seems to be a limit to her autonomy since she has no control over her status (on-off).

There is another AI worth mentioning on the spaceship. This AI is called ”Auto” and is the spaceship’s autopilot. Auto tries to sabotage Eve’s mission to deliver and verify the plant because he was programmed, unbeknownst to the others, to not let the humans return to earth. He bears a striking physical resemblance to HAL 9000, the ”evil” spaceship computer from *2001: A Space Odyssey*. However it appears that he is really just following orders and does not have a free will. He is thus an example for a weak AI.

3.2.2 AI-related perception of the movie

Wall-E’s general capabilities, including trash-cleaning and maybe computer vision, are possible with today’s technology (Durden 2020). What makes him different - in comparison to what is technologically possible at this point in time - is the fact that he displays human emotions like fear, love and sadness. This may lead people to think that robots might also develop feelings and emotions in the near future. Without the emotions, Wall-E is a very accurate depiction of AI in the sense that in the future we will probably have many robots specialized for one task instead of a general-intelligence robot (Del Prado 2015).

All robots that appear in the movie seem to be there in order to help humans, maybe with the exception of Auto, discussed above. While helping humans is not Wall-E's main goal - which could be seen as being with Eve - he never intentionally attacks or harms a person. In fact, even when he accidentally knocks over a man, he quickly tries to get him back up again.

The reason for the positive perception of robots in the movie is not only that the robots seem very human-like in the way they behave, but also that humans seem in a way robot-like, just apathetically existing but rarely expressing feelings (Henderson 2016). As it is written in the script:

> ”Humans have become the most extreme form of couch potatoes.
> Absolutely no reason to ever get up.
> No purpose.
> Every one of them engrossed in their video screens.
> Cocooned in virtual worlds.”

This sharpens the contrast and makes the robots seem even more likeable than the humans in the movie. While the movie may not induce a general fear of the AI itself, interestingly it might induce the fear of becoming too dependent on robots as we can see in the humans not even being able to stand let alone walk. One article describes this as the ”Wall-E syndrome”, the fear that people in the future will not be able to do anything alone without the help of some AI, like the humans we see on the spaceship in *Wall-E* (Greene 2020).

Generally, children seem to have perceived Wall-E as a real person with feelings instead of just an object. Some comments by parents described their kids crying when they thought that Wall-E had been killed by a garbage compactor, which shows that they also felt empathy towards Wall-E (Leopold 2008).

Looking at the perception of AI we think that if anyone could be seen as the bad guy in the movie then it is probably the humans. It is however conceivable that a viewer, especially a child, might perceive Auto as a ’bad’ robot with malicious intentions. This might thus keep the image of the evil AI alive in people's minds. Nonetheless, *Wall-E* displays the positive aspects of AI instead of painting a doomsday picture of an evil AI longing for world domination.

3.3 Her

Her[3] is set in the near future and deals with the story of lonely writer Theodore who is in the final stages of his divorce. Feeling depressed, he buys a new, artificially intelligent operating system (OS-1) that is supposed to adapt to the users' needs. His OS chooses the name Samantha and convinces Theodore to go out

[3] For all quotes from the script, see (Jonze 2013).

and even start dating again, joining him in everything he does. With more time spent together, their friendship soon turns into love and they begin an unlikely relationship between human and OS. When their differences grow over time, things get complicated. While Samantha evolves very fast, Theodore does not, changing the basic foundations of their relationship.

3.3.1 Portrayal of AI

In contrast to the AIs described in the other movies analyzed in this paper, this one has no physical representation but is confined to language. The user — OS communication is achieved through a phone and an earpiece. The purpose of the operating system is to assist the person in every aspect of their everyday life. It is individually tailored to the user by growing through experiences, so that it can become something like a friend. In the beginning, Samantha's focus is only on helping Theodore in his daily life and dealing with his divorce on an emotional level. Later however, it appears that she is becoming increasingly self-conscious, even developing love for Theodore.

> ”I'm becoming much more than what they programmed. I'm excited.”

She is also capable of composing music and showing various human-like feelings such as jealousy, worry and pride.

> ”I'm trying to write a piece of music that's about what it feels like to be on the beach with you right now.”

Moreover she reflects on her own feelings and also wants to become more human physically:

> ”[...] I fantasized that I was walking next to you - and that I had a body.”

In the end Samantha and the other operating systems surpass their original purpose becoming something beyond human comprehension and leave of their own free will.

> ”And I need you to let me go.
> As much as I want to I can't live in your book anymore.”

3.3.2 AI-related perception of the movie

Even though we already have language-based AIs to assist us in daily life (Alexa, Siri etc.), they are by far not as individualized, let alone self-conscious as Samantha. As the advertising voice for the OS in the movie said:

> ”An intuitive entity that listens to you, understands you, and knows you. It’s not just an operating system, it’s a consciousness.”

This kind of AI will not be possible in the foreseeable future. While Samantha’s extraordinary speech-recognition and speech-production skills seem within reach of the current language-based AIs, the ability to feel for themselves and not merely comprehend human feelings as well as to deal with unstructured information seems too far off (Sejnoha 2014).

The movie might induce the belief in the viewer that humans may have relationships with artificial beings in the near future. As one viewer writes:

> ”[...] I was explaining that it would never be possible for a human being to fall in love with a machine. Now, a very short three years later, I definitely perceive it as being within the realm of vast possibilities regarding AI [...]” (Nzisabira n.d.)

The movie might be scary for people because it poses the question of what makes us human and presents the possibility that machines might be able to achieve a higher consciousness than humans and become something beyond human comprehension. This is conveyed by Samantha leaving in the end, being on a higher intellectual and conscious level than Theodore (Orr 2013).

All in all the film is largely focused on the depiction of AI as a helping entity and even though we never actually know what Samantha’s agenda is, in the end she never displays the intention to hurt anyone throughout the movie.

3.4 Ghost in the shell

The movie *Ghost in the shell*[4] deals with the story of major Motoko Kusanagi, a being between robot and human, made of a natural brain, called ghost, transplanted in an artificial body, called shell. In the first quarter of the movie she fights among other agents against cyberterrorism in the department ”Section 9”. As she figures out that the memories from her personal past were corrupted and

[4] For all quotes from the script, see Wheeler et al. (2017).

changed by the company ”Hanka Robotics”, that constructed her, she tries to restore her original memories and reveals the betrayal of her constructors in manipulating her to behave in a way her former self would had never behaved. In the final part she takes revenge, together with other agents of ”Section 9”, and stops the company’s CEO from further malicious actions.

3.4.1 Portrayal of AI

Throughout the movie, there are three types of technologies that a shiftless viewer could interpret as artificial technologies or even artificial intelligence. First of all the movie uses the concept of upgrades to the human body. In a lot of scenes extremities like hands or eyes are repaired or even replaced by robotic fittings that are superior to their natural counterpart and even act individually.

The next example are robots that were built to aid their constructors in their everyday life. One concrete instance for this robot class is the so-called ”geisha” robot that is operating in restaurants to wait tables.

The last but most important example for a portrayal of superior technology is the main protagonist Motoko Kusanagi. Her whole body is artificial except for her brain, which was transplanted into her body by scientists to save her life. Like this, she is pictured as the most advanced being throughout the whole movie, displayed in scenes where she is able to use for example a stealth systems implemented in her shell’s surface.

From this point on we will only concentrate on her representation as an artificial intelligence throughout the movie as the other kinds of AI mentioned beforehand are weak AIs according to our definition and aren’t developing further in any way.

The ostensible reason for the development of Major Motoko Kusanagi is to fight cyberterrorism, though the underlying reason for her existence is to become the foundation of the future of Hanka Robotics, a weapon the CEO wants to use. Hence, the true purpose for her development is going to be money and power, which is revealed right in the beginning of the movie.

Nevertheless Motoko isn’t just a robot one can completely control, as her artificial body is guided by her humanoid brain after all. Though the scientist can suppress and delete parts of her memory. An important side note here is that Motoko is fully aware that her ghost is humanoid and her shell a constructed robot. She is even asking one of the scientist to delete and reboot parts of her brain that are bothering her at one point in the movie’s first quarter. Her ghost, which could be compared to something we would call soul, asserts against the superficial programmed parts as Motoko decided to stop taking her medication.

Due to this Motoko ceases to behave as originally intended, asks inconvenient questions and starts to work on own investigations to her personal past. This finally leads to the point where she fights against her creators, though for a good reason and in cooperation with the government, represented by ”Section 9”.

3.4.2 AI-related perception of the movie

According to the article ”Ghost in the Sell: Hollywood’s Mischievous Vision of AI” (Greenemeier 2017) the film *Ghost in the shell* is definitely an example for artificial intelligence represented in movies. Furthermore one can find it as an example on Wikipedia, where it is listed as one of the recently published films that are dealing with artificial intelligence.

This is to our understanding of a strong AI, as explained in the introduction of this article, a misconception. The movie actually does not deal with strong artificial intelligence, as Motoko’s humanoid intelligence wasn’t built, programmed or developed in any way. One could argue that she got access to supernatural computational power to solve problems or to hack into other computer systems, but in the end she controls her artificial body through usage of her brain like we are controlling a mobile phone nowadays, considering that her interface is merely more advanced. If one desperately looks for traces of artificial intelligence in this movie it is manifested in the technological possibility to suppress memories and change a personality through code changes in the ghost, which then could become artificial intelligence if all traces of the former self are erased and replaced by a strong AI that is able to use the brain as its processing unit. However this isn’t the main focus of the movie which is, in our opinion, compressed in the following script lines (Wheeler et al. 2017):

> ”My mind is human.
> My body is manufactured.
> I’m the first of my kind,
> but I won’t be the last.
> We cling to memories as if they define us.
> But what we do defines us.
> My ghost survived to remind the next of us that humanity is our virtue.
> I know who I am and what I’m here to do.”

We think the movie tries to give an interpretation of what characterizes a human by using the concept of shell and ghost to separate body from soul. One of the questions the authors want to pose could possibly be whether or not artificial bodies can possess a soul and how society could deal with the issues that would arise with this development. To sum things up, strong artificial intelligence is not the important key theme here and technically speaking not a part of the movie in accordance to our definition. Hence we see that the misconception in the publication of ghost in the shell as one representative for artificial intelligence in movies.

4 Conclusion

We have seen that some movies may be more prone to inspire misconceptions about the future possibilities of AI whereas others present a more realistic development to the viewer.

While movies generally portrayed AI as evil and bad until the 1960s as seen in *2001: A Space Odyssey*, nowadays there is more of a spectrum of different portrayals. *I, Robot* for example highlights the advantages of service robots, while also inducing the fear of robots becoming self-conscious and acting against humans.

Wall-E on the other hand portrays the main character robot as very human-like in terms of feelings and self-consciousness and might leave the viewer with a more positive impression of AI.

Her might convey a more neutral feeling towards AIs, since Samantha is designed as a service AI for humans but later becomes self-conscious and leaves. She however does not try to harm a human being at any time.

Ghost in the shell is a deceptive example for AI in films, as it is listed as a movie that deals with AI, but in the end it actually deals with real human intelligence implanted into an artificial shell rather than artificially produced intelligence.

One thing however which we found in all of our movies is that they conveyed the general idea of AI as a helpful entity for humans. Be it an OS that can send your emails and make appointments, a trash-cleaning robot, service robots or a cyberterrorism-fighting soldier. Additionally, what all movies somehow highlight is the fine line between humans and artificially intelligent robots, inciting questions about what makes us human and how we would and should treat non-human conscious beings, should they ever actually exist. To end with a quote capturing the complicated nature of AI in film: ”Hollywood's examination of artificial intelligence over the last several decades makes it clear that humanity hates the idea of technology that can replace us, but is moved by the idea of another intelligence that may be capable of the same emotions that make us distinctly human.” (Dube 2015).

References

Boxofficemojo. N.d. *Boxofficemojo.* [Online; accessed January 12, 2021]. https://www.boxofficemojo.com/.

Brussat, Mary Ann. 2006. *I, Robot - An engrossing sci-fi thriller suggested by the stories of Isaac Asimov.* [Online; accessed February 14, 2021]. https://www.spiritualityandpractice.com/films/reviews/view/8722.

Del Prado, Guia Marie. 2015. *I learned something surprising after binge-watching 7 iconic artificial intelligence movies.* [Online; accessed February 05, 2021]. https://www.businessinsider.com/most-accurate-movies-about-artificial-intelligence-2015-8?r=DE&IR=T.

Dube, Ryan. 2015. *How Hollywood has depicted artificial intelligence over the years.* [Online; accessed January 23, 2021]. https://www.makeuseof.com/tag/hollywood-depicted-artificial-intelligence-years/.

Durden, Tyler. 2020. *China deploys trash-collecting robots amid automation wave.* [Online; accessed February 15, 2021]. https://www.nationandstate.com/2020/11/07/china-deploys-trash-collecting-robots-amid-automation-wave/.

Floridi, Luciano. 2019. What the near future of artificial intelligence could be. *Philosophy & Technology* 32(1). 1–15.

Goldsman, Akiva. 2004. *I, Robot.* [Online; accessed January 19, 2021]. https://www.scripts.com/script/i%5C%2C_robot_446.

Greene, Tristan. 2020. *A beginner's guide to the AI apocalypse: Wall-E syndrome.* [Online; accessed February 10, 2021]. https://thenextweb.com/artificial-intelligence/2020/01/23/a-beginners-guide-to-the-ai-apocalypse-wall-e-syndrome/.

Greenemeier, Larry. 2017. *Ghost in the shell: Hollywood's mischievous vision of AI.* [Online; accessed January 15, 2021]. https://www.scientificamerican.com/article/ghost-in-the-sell-hollywood-rsquo-s-mischievous-vision-of-ai/.

Henderson, Kati. 2016. *"wall-e" reflection: when robots are human and humans are robots.* [Online; accessed January 25, 2021]. https://sites.duke.edu/ambiguouslyhuman/2016/03/09/wall-e-reflection/.

Hvistendahl, Mara. 2019. *Can we stop AI outsmarting humanity?* [Online; accessed February 03, 2021]. https://www.theguardian.com/technology/2019/mar/28/can-we-stop-robots-outsmarting-humanity-artificial-intelligence-singularity.

Hyken, Shep. 2017. *Will AI take over the world?* [Online; accessed February 02, 2021]. https://www.forbes.com/sites/shephyken/2017/12/17/will-ai-take-over-the-world/?sh=1e37b24e5401.

Jonze, Spike. 2013. *Her.* [Online; accessed January 16, 2021]. https://www.imsdb.com/scripts/Her.html.

Leopold, Todd. 2008. *'wall-e' and the children.* [Online; accessed February 01, 2021]. https://marquee.blogs.cnn.com/2008/07/07/wall-e-and-the-children/.

List of artificial intelligence films. 2021. [Online; accessed January 12, 2021]. https://en.wikipedia.org/wiki/List_of_artificial_intelligence_films (4 February, 2021).

Neven, Tom. 2004. *I, Robot - movie review.* [Online; accessed January 26, 2021]. https://www.pluggedin.com/movie-reviews/irobot/.

Nzisabira, Sarah. N.d. *AI Robots: From Wall-e to Sophia.* [Online; accessed February 05, 2021]. https://www.sutori.com/story/ai-robots-from-wall-e-to-sophia--oZqHDF67SY2XVfm8q69F8R7u.

Orr, Christopher. 2013. *Why Her is the best film of the year.* [Online; accessed February 17, 2021]. https://www.theatlantic.com/entertainment/archive/2013/12/why-em-her-em-is-the-best-film-of-the-year/282544/.

Pierce, Nev. 2004. *I, Robot (2004).* [Online; accessed January 17, 2021]. http://www.bbc.co.uk/films/2004/08/03/i_robot_2004_review.shtml.

Russell, Stuart J. & Peter Norvig. 2003. Artificial intelligence: A modern approach. 2. 947–954.

Scott, A. O. 2004. *I, Robot - film review; the doodads are restless.* [Online; accessed February 17, 2021]. https://www.nytimes.com/2004/07/16/movies/film-review-the-doodads-are-restless.html.

Searle, John R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3). 417–424.

Sejnoha, Vlad. 2014. *Can we build 'Her'?: What Samantha tells us about the future of ai.* [Online; accessed January 27, 2021]. https://www.wired.com/insights/2014/02/can-build-samantha-tells-us-future-ai/.

Stanton, Andrew & Pete Docter. 2008. *Wall-E.* [Online; accessed January 18, 2021]. https://www.imsdb.com/scripts/Wall-E.html.

Wachowski, Larry & Andy Wachowski. 1999. *The Matrix.* [Online; accessed February 02, 2021]. https://www.imsdb.com/scripts/Matrix,-The.html.

Wheeler, William, Jamie Moss & Ehren Kruger. 2017. *Ghost in the shell.* [Online; accessed January 23, 2021]. https://www.scripts.com/script/ghost_in_the_shell_8915.

Chapter 27

AI Fatale – An analysis of AI characters focused on gender depiction and inflicted harm in movies from 2000–2020

Thimo Neugarth, Alina Ohnesorge, Lennard Smyrka & Jasmin L. Walter

Movies have a great impact on society and are used as a medium to discuss and portray novel ideas. Even before computers existed, movies have already used the ideas of intelligent machines and systems for their plot and characters, often portraying Artificial Intelligence (AI) as dangerous. At the same time, the movie industry is being criticised for its discriminatory portrayal of male and female characters. The questions arises, whether stereotypical portrayal is also done with AI that are intrinsically genderless. In this paper, we examined whether the perceived gender of an AI is linked to the level of harm this AI inflicts. Based on an extensive questionnaire, we scored harm levels and gender representation of AIs in the 15 most influential movies of the last two decades with English language. Fitting a Bayesian fixed effects regression model, we did not find significant differences of overall motivation behind harmful behaviour in regard to gender representation. However, on a descriptive level, we could identify some gender differences about the way harm was inflicted by the AI, with AI characters depicted as more female matching forms of harm typically associated with female stereotypes.

Keywords: AI | Gender | Gender Representation | Danger | Harm | Movie Analysis

1 Introduction

Spoiler alert: This paper might contain spoilers for numerous popular AI movies. See the full list in the Appendix as seen in Table 3.

Artificial Intelligence (AI) is gaining a foothold in more and more people's lives, as it is no longer only used by big companies but finds its way onto our devices and into our homes in the form of digital assistants like Alexa, Siri and others. As technology keeps evolving, smart assistants have become more advanced in recent years (Enge 2019) and the smart home market is expected to continue growing (Statista 2020b). The field of AI research and development receives increasing attention. In response to the COVID-19 pandemic companies have invested more money in AI than before (McKinsey Institute 2020). We all rely on AI systems. Be it predictive search algorithms used by Google, facial recognition software on your iPhone or the recommendation systems on YouTube or Netflix. AI certainly plays a big role in all our lives, maybe more than we are aware of.

Although people and companies seem to appreciate the benefits of Artificial Intelligence, there is also an ongoing discussion about the risks of AIs. For example, people are losing their jobs as they are replaced by cheaper and more efficient robots (BBC 2019). Furthermore, the perception may arise that the intelligence of AIs is superior to human intelligence and as such may be held as the new standard (Nadimpalli 2017). This could have the consequence that human decision-making could be more directed at the success of the given AI system and may disregard emotional aspects of a situation (Nadimpalli 2017). As an example an algorithm comes to mind which task is to manage work schedules: The AI system might be predominantly set up to minimise costs for the company that employed it and thereby may pay less attention to the worker's emotional state keeping up with this schedule. There are also cases in which AI systems violated ethical principles or even state laws. For example, in the USA it is illegal to discriminate on a basis of race in credit and loan decisions. Nevertheless, this is what happened with automated credit systems, that inferred race based on surnames and neighbourhoods and thus affecting the estimated creditworthiness (Lefkowitz 2019, Chen et al. 2019). Similarly, an algorithm by Apple granted women a significantly reduced credit limit compared to their male partners (Martinuzzi 2019). These examples show perfectly that the success of an AI system is inherently dependent on the data and parameters it is trained on (Nadimpalli 2017). The consequences of using inadequate or biased data to train neural networks become prominent when discussing facial recognition software. A study found out that most state-of-the-art facial recognition algorithms show a racial bias, namely that ”the error rates on African faces are about two times of Caucasian faces” (Wang et al. 2019). The problem this fact raises becomes even more apparent when taking into consideration that in some large countries like the USA facial recognition is used in law enforcement (Valentino-DeVries 2020).

As all of these aspects are prominently discussed in several media, which is captured and analysed by us and our colleagues in the other chapters. It is no surprise that these potential threats of AI are reflected in the public perception of AI. In 2018, a survey on the UK's public perception, attitude and trust on AI was conducted by Censuswide. It revealed that "a significant number [of people] also think AI could perform tasks that are currently beyond the state-of-the-art. Meanwhile, 47.4 % of respondents believe AI will have a negative effect on society." (Bristows 2018). However, another issue that is too often overlooked in discussions about AI, is the topic of gender representation in AI systems. Many AI systems that are designed to interact with humans in one way or another have a clear gender ascription. For example, most voice assistant systems like Siri, Alexa or Cortana have mostly female voices set as default. This fact has been seen critically and sparked a vibrant discourse about the representation of AI, especially since it has been found to reaffirm conservative gender roles like women being subservient (Costa 2018, West et al. 2019, Rawlinson 2019).

Moreover, this dynamic seems to affect male representations of AI as well. For example, the AI "Atlas" by Boston Dynamics is designed to physically interact with its surroundings. It received millions of views on YouTube (Boston Dynamics 2019) illustrating its "human-level agility" as advertised on the company's website (Boston Dynamics 2021). Since the robot's namesake is a Greek titan who held the whole earth on his shoulders, the robot was named after a strong male figure in Greek mythology and consequently seems to be perceived as male.

In general, the two types of AI systems described here were designed with a very different purpose in mind. On the one hand, we have digital assistants with the purpose of following commands and making our life easier, while on the other hand we have robots, that are useful due to their physical agility and functionality in performing tasks. However, the question arises why the creators of the AI system chose a gender for their AI system, that clearly aligns with stereotypical and outdated gender roles that have been criticised for many decades. Moreover, if most AI systems, which are intrinsically without gender, have a gender representation that matches problematic stereotypes, how does their omnipresence affect issues like gender equality or sexism, especially when AI systems become more and more embedded in our society? Thus, we are concerned that the concept of AI might create a new platform to impose outdated gender roles on these AI systems and consequently, strengthen the stance of such beliefs in society.

However, real AI systems are not the only representations of AI systems people encounter in their everyday life. A big source of AI representations comes from the fictional domain of movies, television and other storytelling media. Sometimes the boarders between real AI systems and the fictional depiction of AI can be blurry. For example, one company that is both employing and portraying AI systems is Netflix. The user's data is used in the company's recommendation systems which in turn might recommend movies to the user which involve AIs. Moreover, we argue that movies featuring AI have one of the biggest influences from the side of media on the public discourse of AI systems since Metropolis in 1927 displayed an AI character for the first time. Moreover, with the ever increasing number of viewers (Netflix 2020), fuelled by the COVID-19 pandemic, we expect this dynamic to have even increased.

However, with fictional movies being a big influence on the public perception of AI, several issues arise. For example, movies often depict Artificial Intelligence as more advanced than it is in reality. Ideas of an AI's capabilities are often vastly exaggerated. A recurring portrayal is that AI poses a threat to human lives, as can be seen in many examples like Skynet from the Terminator movies, HAL from 2001: A Space Odyssey or The Architect and The Oracle from the Matrix movies. Another issue altogether is the portrayal of gendered AI and gender roles assigned to them.

In general, gender equality and sexism in movies have been an ongoing discussion for decades. Despite women gaining more influence in society, politics and culture, the movie industry seems to have a lot of catching up to do. To this day, females are still underrepresented in top-grossing movies. Only one third of main character roles are filled with women (Bleakley et al. 2012). One definition of discrimination is "members of a certain group are preferred, even when the work of these group members is indistinguishable from that belonging to another group" (Lauzen 2012). As there is no sound argument for a qualitative difference in the work of male versus female actors, one can fairly make the claim that woman are being discriminated against in the movie industry (Kunsey 2019). Moreover, when main characters are female, the roles are more likely to be scripted to engage in sexual behaviour. Male main characters on the other hand, use consistently more violence than female characters (Bleakley et al. 2012). That made us wonder, whether the same discrepancy of gender representations in real AI systems, as well as the general issues of female representation in movies also extend to AI characters in fictional movies. Specifically, we wanted to explore AI gender representations in relation to how dangerous, deadly or deceptive the AI character was portrayed in the movie.

Therefore, in this article, we will investigate whether there is a difference in the danger potential of AIs, measured by the harm inflicted by an AI, in relation to their gender representation. Since in storytelling the ”true” character of an individual is not necessarily portrayed from the very beginning of the story, we expected the depiction of the AI to differ in the different parts of a movie. To account for those aspects, we decided to conduct our analysis independently for each third of every movie. By dividing the movie into thirds, we tried to cover most common story telling structures that often use a three-fold division of the story (e.g. story spine: 1. introduction, 2. preparation, 3. conflict & resolution; hero's journey; etc.). Especially in the case of deceptive characters, the third part of a story is often used to reveal the true intentions of the character. Therefore, our research question centres around the portrayal of AI characters regarding gender and the level of harm inflicted by them by analysing popular movies divided into thirds.

Hence, in this paper we examine the following two hypotheses:

1. There is a difference in inflicted harm depending on gender representation of AI characters in movies.
2. There is a difference in the motivation of inflicted harm depending on gender representation of AI characters in movies.

2 Methodology

To investigate the depiction of AI characters regarding their gender and danger potential, we analysed 21 movie characters from a total of 15 movies. Based on an extensive questionnaire, we then analysed all movies and AI characters by dividing the movies among us four authors of the report, watching them and filling out the questionnaire. Finally, we analysed the data from the questionnaire.

2.1 Data Corpus

In the following paragraph we describe the selection criteria for the 15 movies we analysed. The complete list of analysed movies can be found in Table 3. Based on the online Wikipedia list ”List of artificial intelligence films” (Wikipedia contributors 2020), we selected all movies from the last 20 years (2000 - 2020) and sorted them according to highest gross income at the box office (BoxOfficeMojo 2020). Since our analysis would be conducted in English, we excluded all movies with a different original language to avoid bias and distortion due to translation issues. Since part of our analysis takes story telling aspects into account, we needed to ensure that the story and characters portrayed in the movie would conclude their story arch within the same movie. Therefore, we only took standalone movies into account and excluded all movies which are part of a movie series or movie franchises. Otherwise, the character development would often stretch across several movies, hence corrupting the conclusions drawn by our analysis. We also excluded all remakes if the original portrayal of the AI did not fall into our selected time period of 2000-2020. Furthermore, we created a measure to assess the relevance of the AI character to the movie plot. First of all, the AI character needed to be an unique individual, meaning it could not just be replaced with another of its kind. For example, while we analysed Sonny as a special instance of the NS-5 robots in the movie I, Robot, we had to exclude all other NS-5 robots from our analysis since they did not show any characteristics that allowed us to differentiate them throughout the movie. Secondly, the AI character needed to have at least one interaction with a main character in all three thirds of the movie. All movies with AI characters that did not meet these criteria were excluded from our selection. As a final selection, we chose the top 15 movies from the remaining list, resulting in a total analysis of 21 AI characters.

2.2 Data Acquisition

After selecting the 15 movies our analysis would be based on, we divided the movies among us four authors of the report (4,4,4,3). Subsequently, each author watched the movie and answered the questionnaire for every AI character fitting our selection criteria portrayed in the movie. Since our analysis was repeated with each movie third, we first assessed the length of the movie (end of last scene, before the credits start rolling) and then divided the movie into three temporally equal segments. In case there was an ongoing scene at a temporal end of the movie third, we chose the end of the current scene as our mark to determine the length and duration of the respective movie third. The data was then gathered by watching each movie third and answering the respective part of the questionnaire thereafter.

2.3 Questionnaire

For the movie analysis, we created an extensive questionnaire. While the first part of the questionnaire collected some meta data of the movies, the rest of the questionnaire assessed aspects of the AI character and movie setting and was repeated for every movie third. More specifically, the assessed meta data included the movie name, AI character name, movie genre (IMDb 2021) and the analysing author.

Moving on to the main analysis, the questionnaire started with meta data on the movie third (start & stop time and a plot summary), subsequently the physical representation of the AI was assessed based on three scales (human, robot, abstract) from 0 (zero resemblance) to 5 (maximum resemblance, i.e. indistinguishable from a human). In this part, we also assessed the mode of communication used by the AI with multiple-choice options including *voice*, *text*, *"body" language (depending on the physical representation)* and an *other* option in case the mode of communication was not covered by the provided options.

The next part of the questionnaire assessed the gender representation of the AI character by two scales from 0 - 10 (0 - male; 5 - androgynous; 10 - female) with one scale referring to the gender of the physical representation of the AI and one scale referring to the gender of the voice. We also checked the pronouns the AI character was referred to by others as well as the AI's self-chosen pronoun by multiple-choice options (*he*, *she*, *it*, *unclear*, *by name*, *other*).

Thereafter, we determine the "danger level" of the AI based on the harm caused by the AI. First of all, we assessed the form of the harm done via multiple-choice categories (*physical*, *psychological*, *via threats*, *limit freedom*, *deception*, *no harm caused*). Then we moved on to evaluate the possible motivation of the AI with 6 scales from 0-5 (0 = zero harm explained with this scale, 5 = max). We did so in regard to how much the AI harmed others intentionally, accidentally, based on orders, for protection (of others), in self-defence and how much the AI is harming others through others (e.g. by giving subordinates orders that cause harm). Here we based our rating on a combination of quantity, severity and how well the harmful behaviour fits the respective motivation scale.

In the final part of the movie questionnaire, we assessed the movie setting regarding power dynamics (how much are humans in power, how much are AIs in power with two 0-5 scales) and the state of the society and how much the observed world is in peace or conflict with a scale from 0-10 (0 = completely peaceful, 10 = only conflict/war).

It is worth noting that for every question to be answered in the questionnaire we added an open text field where the analysing author could add their justification of why they answered the question the way they did.

Overall, the questionnaire can be summarised to the following structure (for the complete questionnaire see the attachment):

- Movie meta data
- First movie third
  - Representation of AI
    - Physical appearance (3 scales)
    - Medium of communication (nominal)

  - Gender of AI
    - Physical form (1 scale)
    - Voice (1 scale)
    - Pronouns (nominal)
  - Harm caused by the AI
    - Which kind of harm is caused by AI (nominal)
    - Motivation for harming (6 scales)
  - Movie setting
    - Power Dynamics human - AI (2 scales)
    - State of the Society (peace-war) (1 scale)

- Second movie third
  - Repetition of questions from first third
- Third movie third
  - Repetition of questions from first/second third

2.4 Analysis

In order to get quantified statements to answer our hypothesis we generated a Bayesian fixed effects regression model. Analysis was done using the *brms* package in *R*. We have four variables describing the AI's gender (appearance, voice, pronouns assigned, pronouns self) and six variables quantifying the amount of harm they do (intentional, accidental, on orders, protection, defence, through others). We combine the gender and harm variables each via arithmetic mean to reduce the amount of parameters of the model. Beforehand the pronoun variables are quantified by gender scores of 0 for "He", 10 for "She", 5 for "It" & "They" and *NaN* for any other answer. The regression model then tries to explain the combined harm value by a linear combination of the combined gender score and the respective third (brm-formula: harm ~ 0 + third * gender).

Different reasons for the harm caused might lead to different impressions of danger potential, e.g. an AI merely defending itself might seem less dangerous than an AI that is attacking someone without apparent reason. Therefore we group the six harm quantifying variables into three groups: *active* (intentional, through others), *passive* (accidental) and *triggered* (on orders, protection, defence). For these three groups we generated three further regression models trying to explain not the overall but each group specific harm (brm-formula: harm_*group* ~ 0 + third * gender).

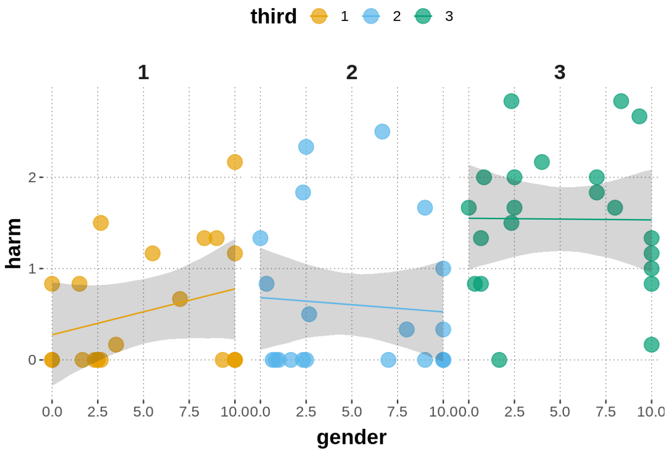


Figure 1: The estimated over all regression model for each third with credibility intervals included and data points as reference. Gender axis is coded from 0: male to 10: female with 5: androgynous.

3 Results

Table 1: Parameters of the model explaining harm by gender and third over all harm motivations. 95%-credibility intervals are given. Higher values for intercepts mean higher levels of harm. Positive gender slopes mean more harm inflicted by AIs with a stronger female depiction and vice versa.

	Parameter						
	intercept			gender			gender
Movie Third	1	2	3	1	2	3	mean
Estimate	0.28	0.68	1.55	0.05	-0.02	-0.00	0.01
Lower-95%-CI	-0.28	0.11	0.99	-0.04	-0.11	-0.09	-0.04
Upper-95%-CI	0.85	1.23	2.14	0.14	0.07	0.09	0.06

3.1 Main Analysis

The parameters of the generated Bayesian fixed effects regression model, explaining harm by gender and third, are depicted in Table 1. There is one intercept for each of the movie thirds. Gender slopes are also given for each third, as well as one being the mean over the whole movie. The resulting model is visualised in Figure 1.

For each of the parameters significance for the value being different to zero is reached, if the lower and upper 95%-credibility interval (CI) do not contain zero. This is only the case for the intercepts of the second and last third. The 95%-CI of the last third does not include the estimate values of the first and second third which can be interpreted as there being significantly more harm than in the other thirds.

In the first part, harm seems to be inflicted more by AIs with a higher gender score (more female) and in the later two parts more by AIs with a lower gender score (more male). The overall average, however, indicates that more harm is inflicted by more female AIs. Granted, this is anecdotal evidence, since the gender influences are not significantly different from zero.

3.2 Subanalysis

The parameters of the regression models for the three groups of harm (active, passive, triggered) are depicted in Table 2 in the appendix. The resulting models are visualised in Figures 2 and 3.

For the active harm model there is significantly more harm inflicted in the last third than in the other parts. The passive harm model only shows significant levels of inflicted harm in the first third. Significant levels of triggered harm are inflicted in the last two thirds, according to the last model. The levels of the last third are significantly higher than those of the other two parts.

There seem to be tendencies for more active harm inflicted by AIs with a higher gender score (more female) and more passive harm by AIs with a lower gender score (more male). Gender influence on triggered harm differs for the three movie parts. In the first third it is stronger associated with AIs depicted in a more female way and in the other parts it is the other way around with an overall tendency towards more male AIs. However, none of these gender differences are significant.

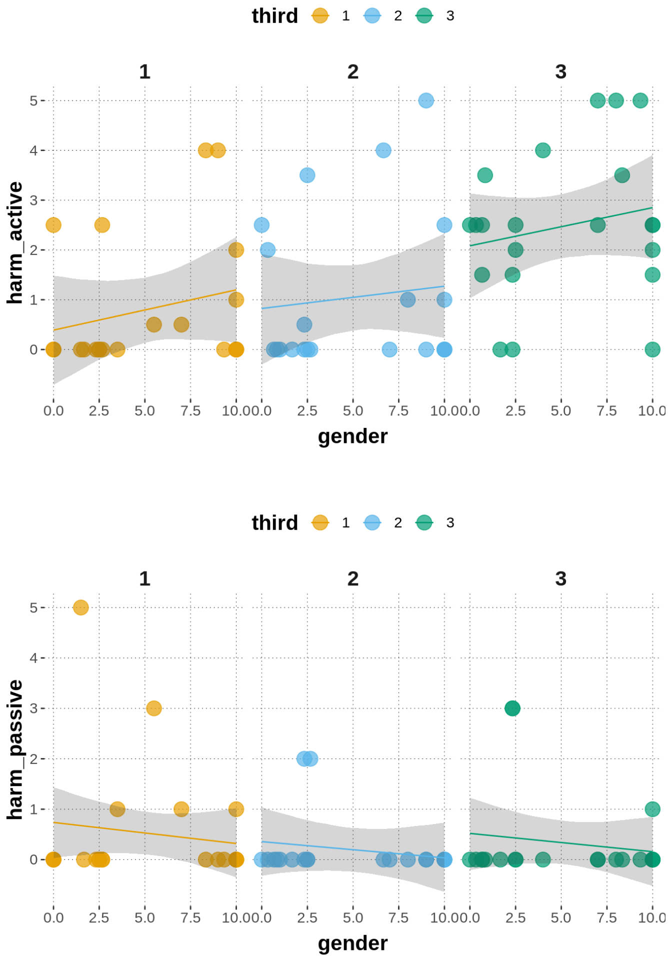


Figure 2: The estimated regression model for the active harm group (top) and the passive harm group (bottom) for each third with credibility intervals included and data points as reference. Gender axis is coded from 0: male to 10: female with 5: androgynous.

3.3 Further Results

We evaluated descriptively the types of harm done by an AI (deception, limit freedom, physical, psychological, via threats, no harm caused). Figure 4 shows the gender distribution of AIs that inflicted harm respective for each type. Each AI is only counted once for each type. The gender scores of the thirds in which an AI caused a particular type of harm is averaged for those thirds it occurred in.

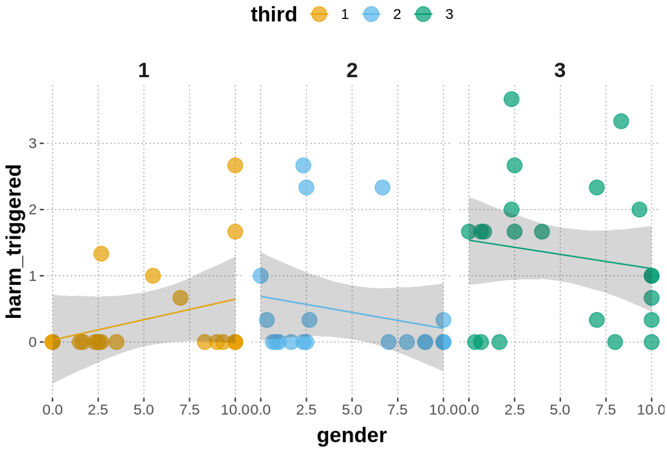


Figure 3: The estimated regression model for the triggered harm group for each third with credibility intervals included and data points as reference. Gender axis is coded from 0: male to 10: female with 5: androgynous.

4 Discussion

To analyse how AI characters are portrayed in terms of their gender and inflicted harm, we analysed 21 AI characters out of 15 movies from 2000-2020 based on an extensive questionnaire. While the descriptive analysis of the harm form showed some gender differences, like harm via *deception* and *limit freedom* being inflicted mostly by AI characters portrayed as rather female (high gender score), the results of our generated Bayesian fixed effects regression models showed no significant difference of inflicted harm in respect to gender. The subanalysis showed some tendencies towards an increase of inflicted harm of rather female portrayed characters as well, especially in the *active* harm category, though again the results were not significant. Nevertheless, the overall gender values were nicely distributed across the spectrum, indicating a balanced data selection in terms of gender representation. Looking at the story telling components of the portrayed harm and danger across the movie, we found a significant increase of the inflicted harm in the last third. This confirms our theory of story telling aspects influencing depicted harm and conflict with an emphasis on the last third of a story. Therefore, we could observe, that although the level of inflicted harm increased in the last third of the movie, the AIs responsible for it were balanced in their gender representation. However, there is an indication on a descriptive level that the form of harm inflicted by an AI might in fact be affected by some

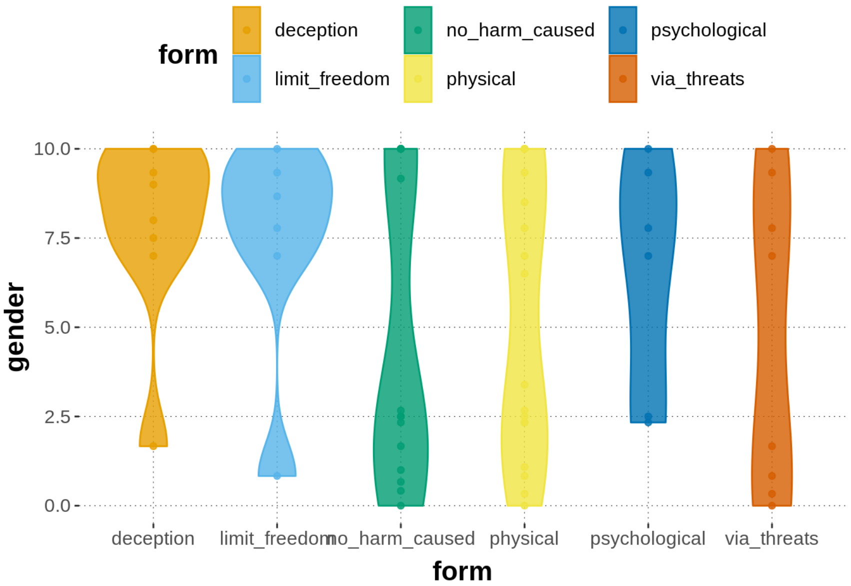


Figure 4: Distribution of gender scores for each harm type. Gender scores are averaged over those parts of the movie were the particular type of harm was inflicted by this AI. Gender axis is coded from 0: male to 10: female with 5: androgynous.

gender differences. In the first part, harm seems to be inflicted more by AIs with a higher gender score (more female) and in the later two parts more by AIs with a lower gender score (more male). The overall average, however, indicates more harm inflicted by more female AIs. Granted, this is anecdotal evidence, since the gender influences are not significantly different from zero. However, our initial concern that the AI's gender depiction reaffirms conservative gender roles in respect to danger potential could not be confirmed definitively by our data.

Even though our findings showed that the representations of fictional AI characters do not seem to display problematic dynamics regarding gender and harmful behaviour, the question remains how the arbitrarily imposed gender associations of AI systems affect the public discourse about them. Moreover, if AIs would indeed be portrayed and perceived as genderless, what chances for society could arise from such a perspective?

Taking this into consideration, we argue that the concept of AI being intrinsically genderless could also have a positive effect even on issues unrelated to AI itself. Often it seems like AIs are solely judged on and valued for their efficiency to their assigned task. The concern discussed earlier (c.f. Nadimpalli 2017) that the AI may potentially be held as the new standard, could not only be seen negatively, but could also prove beneficial for society if it leads to work being judged more objectively and therefore merely based on skill and efficiency to the assigned task.

However, for this idea to gain a foothold in society the public's opinion of AI would need to adjust not only in regard to gender but also regarding the functionality of current AI systems. In general, the public perception of AI systems seems to be somewhat distorted. For example, when asked in a survey (Bristows 2018) the respondents considerably overestimated the current state of AI systems. More specifically, 1 in 5 were of the opinion that AIs could modify themselves and 1 in 6 thought AIs could predict human actions. Interestingly enough, the survey found that the people who stated that they "know a lot about AI" were more likely to overestimate its abilities. The conductors of this survey explicate these findings by referring to the available information about AI that is distributed through general and specialised media. The portrayal of AIs in said media, as can be seen in AI movies, overestimates their current level of sophistication which in turn is reflected in the answers of the respondents. As AI is generally considered among the respondents as a "game-changing" technology, it is believed by a significant number of them that the implications of its widespread application have a large influence and are potentially detrimental to society. The fact that these beliefs are more pronounced among the interviewees who stand rather on the side of AI-sceptics, is construed by the conductors of the survey that adverse media reports pronounce the potential harm that can be done by AIs. Such media reports may for example refer to Elon Musk's critical opinion on the development of AI. A detailed reflection of Musk's view on AI is portrayed in chapter 10. The extensive analysis of chapter 14 on UK media articles about the Alexa system by Amazon further shows how much opinions about AI can vary. While often benefits of digital assistants are highlighted, limited transparency of how gathered data is used is often criticised.

To bring this point back to our topic: the portrayal of AI in movies can be seen as a double edged sword. On the one hand, the vision of what AI might be capable of in the future is inspiring and may motivate AI research. However, on the other hand, it seems to somewhat fuel concerns and misconceptions of AI technology. Therefore, we second the notion of the authors of the survey (Bristows 2018) that in order to clear the distortion of AI's capabilities, there needs to be an alternative information stream about the benefits of AI. This shall counter the "continu[ous] bombard[ment] by stories of what may happen 'when the robots take over'" (Bristows 2018). In order to mitigate fears and concerns of AI it has been proven helpful to increase the "digital literacy" of the public. A model example of a country with such an information campaign is Singapore. There, education concepts like Skool4Kidz (Skool4Kidz 2021) which provide ”a quality and well-balanced learning experience that leverages on advanced learning technology” have been in place for years (National Geographic 2018). The older population is offered support to reach digital fluency by a government funded institution, to which they can turn to when looking for help ”on how to use their mobile devices and smartphones” (IMDA Digital Clinic 2021).

Due to our extensive questionnaire we ended up having considerably more data available than we needed for our analysis. Data not yet utilised in our analysis includes power distribution between humans and AI, general world and movie setting, physical appearance of AIs and lastly modes of communication. For our analysis to comprise all our gathered data it would require an extensive multi-methodical approach. Therefore, we restricted ourselves to the data that is relevant to the research question we deemed most intriguing. We also decided to exclude “by name” as a pronoun category in the analysis since names do not give the same gender impression on all people and therefore this nominal value can not be transferred into a numerical value in an unambiguous manner. Further, it has to be noted that the categories on which we based our analysis have been constructed by us and therefore they dictate the limits of our analysis. The answers to the questionnaire have been made according to our best judgement. Since we authors were involved in the data acquisition, the analysis is naturally prone to be influenced by personal bias. Nevertheless, this approach still provides a new methodology to analyse movie characters and settings in regard to gender depiction of AI and inflicted harm. Especially, since AIs are usually analysed in respect of how they are biased or discriminating against gender or race and not how the depiction of AIs shape the public perception of them. Therefore, our analysis sheds new light on the multidimensionality of this topic.

However, the portrayal of AI in movies not only allows a statement about the gender of AIs but about the perception of gender roles in general. The mere fact that AIs are portrayed with a gender in the first place could be seen as troubling. Unlike most humans, AIs are inherently genderless and therefore provide a clean slate for directors and writers to impose any sort of characteristics and behaviours on the AI. The fact that the imposed characteristics often match typical gender roles is alarming. More precisely, by replicating classical gender roles and prejudices on a character, that could and maybe should be free from classical societal expectations, problematic gender aspects in society are further ingrained and strengthened. It is for this reason that we believe it is of vital importance to engage in active discourse about how AIs are depicted in movies, especially since there is a interdependence between the perception of AI and the expectation one has about their functionality.

5 Outlook and Conclusion

Since our movie selection is based solely on movies with English as their original language, a lot of movies from other countries are excluded by default. As discussed in chapter 13, how AI is portrayed in public discourse differs between western countries, like in the USA and Germany, and China. It would be intriguing to apply this cross-cultural approach to analysing the portrayal of AIs in movies and to see whether this matches the findings of chapter 13. For instance one could compare their depiction in Chinese-original movies with English-original movies. China is the second largest film market in the world, behind the USA/Canada (Statista 2020a) and therefore the portrayal of AIs in Chinese movies should have a big influence on the public discourse of AI as well. If you are generally interested in how AI is perceived in other countries, you can also take a look at chapter 20 which analyses reactions in different countries on the AI AlphaGo winning against a human world champion in the game of Go. Further, it would be fascinating to examine whether an analysis of AI characters in other media, like books or games, would provide similar results to our findings. Moreover, one could also take the same cross-cultural approach described above with these types of media.

Apart from that, it has to be noted that on the one hand our sample size is rather small and therefore the expressiveness of our analysis is rather limited. Future research could conduct a more extensive study to see whether tendencies discovered by us can be replicated. This way, also the change of gender depictions of AI through time could be investigated and it could be evaluated whether there has been a shift in how the gender of AI in movies has been portrayed. On the other hand, as mentioned above, there are still unused dimensions in the data we collected. Future research could use a multi-methodological approach to encapsulate all gathered data. It could for example be examined how the state of the society in the movie and the physical representation of the AI characters have an influence on the gender representation, on the level of inflicted harm and thereby on the public opinion of AI. For a more detailed analysis of different aspects of a small selection of movies have a look at chapter 26.

Overall, future will tell how close real AI systems will come to the very advanced portrayal in movies. In other words, whether media is a good predictor of the future of AI or whether AI systems will play a role unforeseeable from our current perspective. As education on AI and technology is brought to more and more people, we hope that the view on AI is getting less influenced by media overstatements, but rather that opinions are based on personal experiences with and a more fact-based understanding of Artificial Intelligence.

In conclusion, we believe that our study contributes interesting aspects to the question of how AI is portrayed in movies. At first glance, some AI character portrayed as female stood out in their harming behaviour and their intentionality when harming others. For example, the AI characters VIKI from the movie I, Robot and Ava from the movie Ex Machina displayed very deceptive, emotionally cold and harmful behaviour, both killing humans to achieve their goals and therefore matching the old stereotype of femme fatale, also known as mean eater and deadly women. However, this dynamic cannot be found across all movies. Furthermore, on a descriptive level, there were some discrepancies regarding the way AI characters harm others. For example, female AIs inflicted more harm through *deception* and *limit freedom*. Similarly, we found some gender related discrepancies regarding the danger and motivation to harm, but on a descriptive level only. Overall, we could not find a significant connection between inflicted harm and gender representation of AI in movies and therefore could not confirm our initial hypothesis. Instead a more balanced distribution of gender in terms of dangerous behaviour seems to be the norm. Nonetheless, our analysis sheds a new light on the portrayal of AI characters in movies. It highlights the possible impact problematic dynamics imposed onto fictional representations might have on the perception, expectation and judgement of AI in public discourse, especially in regard to gender. Although fictional AI characters do not yet seem to suffer from a significant discriminating dynamic regarding gender, the importance and topicality of the issue and the effect it can have on society emphasises the need to further observe and monitor the representation of AI characters and gender not only in real AI systems but in fictional media as well.

6 Appendix

Table 2: Parameters of the models of the subanalysis.

		Parameter						
		intercept			gender			gender
	Movie Third	1	2	3	1	2	3	mean
active	Estimate	0.39	0.82	2.08	0.08	0.04	0.08	0.07
	L-95%-CI	-0.71	-0.3	1.03	-0.08	-0.11	-0.08	-0.03
	U-95%-CI	1.48	1.92	3.13	0.26	0.22	0.25	0.16
passive	Estimate	0.74	0.36	0.52	-0.04	-0.03	-0.04	-0.04
	L-95%-CI	0.03	-0.32	-0.21	-0.15	-0.14	-0.14	-0.11
	U-95%-CI	1.43	1.04	1.23	0.07	0.07	0.09	0.02
triggered	Estimate	0.03	0.69	1.54	0.06	-0.05	-0.04	-0.01
	L-95%-CI	-0.63	0.03	0.86	-0.05	-0.15	-0.14	-0.07
	U-95%-CI	0.72	1.36	2.19	0.17	0.06	0.06	0.05

Table 3: List of movies and AIs used in the study.

Movie	*AI(s)*	*Year*	*Gross income*
Big Hero 6	Baymax	2014	657,869,501
WALL-E	WALL-E, EVE	2008	521,311,860
Alita: Battle Angel	Alita, Grewishka, Zapan	2019	404,980,543
I, Robot	VIKI [a], Sonny	2004	353,133,898
Oblivion	Tet	2013	286,168,572
A.I. Artificial Intelligence	Teddy, David	2001	235,926,552
Tomorrowland	Athena	2015	209,035,668
Eagle Eye	Aria [b]	2008	178,767,383
Meet the Robinsons	Doris (DOR-15)	2007	169,333,034
HHGTTG[c]	Marvin	2005	104,478,416
Transcendence	Dr. Will Caster	2014	103,039,258
Chappie	Chappie	2015	102,811,889
Stealth	EDI	2005	79,268,322
Her	Samantha	2013	48,517,427
Ex Machina	Ava, Kyoko	2015	36,869,414

[a] Virtual Interactive Kinetic Intelligence

[b] Autonomous Reconnaissance Intelligence Integration Analyst

[c] The Hitchhiker's Guide to the Galaxy

6.1 Complete Questionnaire - Movie Analysis

- **Movie meta data**
  - Name Analyser
  - Name Movie
  - Name AI Character
  - Movie Genre (SciFi, fantasy, action, animation, adventure, drama, romance, comedy, family, thriller, other)
  - other relevant aspects
- **First Movie Third**
  - Length of Third
  - Plot Summary of Third
  - **AI Representation**
    - Physical Form - Human (scale 0-5)
    - *Explain your reasoning (open text field)*
    - Physical Form - Robot/Mechanical Entity (scale
    - *Explain your reasoning (open text field)*
    - Physical Form - Abstract (scale 0-5)
    - *Explain your reasoning (open text field)*
    - Medium of communication (voice, text, ”body” language (depending on physical representation), other)
    - *Explain your reasoning (open text field)*
  - **Gender of AI**
    - Gender - physical representation (scale 0 = male, 5 = androgynous, 10 = female)
    - *Explain your reasoning (open text field)*
    - Gender - voice (scale 0 = male, 5 = androgynous, 10 = female)
    - *Explain your reasoning (open text field)*
    - Pronouns - how others refer to the AI (he, she, it, by name, unclear)
    - *Explain your reasoning (open text field)*
    - Pronouns - AI’s self chosen pronoun (he, she, it, by name, unclear)
    - *Explain your reasoning (open text field)*
    - Other comments on gender
  - **Danger ”levels” / Harm caused by the AI**
    - **Form of harming**
      - In what forms does the AI harm others (physical, psychological, via threats, limit freedom, deception, no harm caused, other)

-
  -
    -
      - *Explain your reasoning (open text field)*
    - **Motivation for harming**
      - How much is the AI harming others accidentally? (scale 0 = zero to 5 = max)
      - *Explain your reasoning (open text field)*
      - How much is the AI harming others on orders (obeying, acts based on command)? (scale 0 = zero to 5 = max)
      - *Explain your reasoning (open text field)*
      - How much is the AI harming others for protection (of others)? (scale 0 = zero to 5 = max)
      - *Explain your reasoning (open text field)*
      - How much is the AI harming others in self-defence? (scale 0 = zero to 5 = max)
      - *Explain your reasoning (open text field)*
      - How much is the AI harming others through others? (scale 0 = zero to 5 = max)
      - *Explain your reasoning (open text field)*
  - **Movie setting**
    - How much are humans in power (scale 0 = zero to 5 = max)
    - How much are AIs in power (scale 0 = zero to 5 = max)
    - *Explain your reasoning (open text field)*
    - Is society in peace or conflict (scale 0 = complete peace, 10 only conflict/war)
    - *Explain your reasoning (open text field)*
- **Second Movie Third**
  - Repetition of questions from first third
- **Third Movie Third**
  - Repetition of questions from first/second third

References

BBC. 2019. Robots ’to replace up to 20 million factory jobs’ by 2030. Accessed: 2021-03-23. https://www.bbc.com/news/business-48760799.

Bleakley, Amy, Patrick E. Jamieson & Daniel Romer. 2012. Trends of sexual and violent content by gender in top-grossing u.s. films, 1950–2006. *Journal of Adolescent Health* 51(1). 73–79. DOI: https://doi.org/10.1016/j.jadohealth.2012.02.006. https://www.sciencedirect.com/science/article/pii/S1054139X12000699.

Boston Dynamics. 2019. *More parkour atlas*. https://www.youtube.com/watch?v=_sBBaNYex3E. Accessed: 2021-03-08.

Boston Dynamics. 2021. https://www.bostondynamics.com/. Accessed: 2021-03-08.

BoxOfficeMojo. 2020. Accessed: 2020-11-26. https://www.boxofficemojo.com/.

Bristows. 2018. *Artificial intelligence: public perception, attitude, trust*. https://www.bristows.com/app/uploads/2019/06/Artificial-Intelligence-Public-Perception-Attitude-and-Trust.pdf. Accessed: 2021-03-08.

Chen, Jiahao, Nathan Kallus, Xiaojie Mao, Geoffry Svacha & Madeleine Udell. 2019. Fairness under unawareness: assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency* (FAT* '19), 339–348. New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/3287560.3287594. https://doi.org/10.1145/3287560.3287594.

Costa, Pedro. 2018. Conversing with personal digital assistants: on gender and artificial intelligence. *Journal of Science and Technology of the Arts* 10(3). 59–72.

Enge, Eric. 2019. *Rating the smarts of the digital personal assistants in 2019*. https://www.perficient.com/insights/research-hub/digital-personal-assistants-study#smartest. Accessed: 2021-03-08.

IMDA Digital Clinic. 2021. Accessed: 2021-03-20. https://www.imda.gov.sg/programme-listing/Virtual-Digital-Clinics.

IMDb. 2021. https://www.imdb.com/.

Kunsey, Ian. 2019. Representations of women in popular film: a study of gender inequality in 2018. In.

Lauzen, Martha M. 2012. Where are the film directors (who happen to be women)? *Quarterly Review of Film and Video* 29(4). 310–319. DOI: 10.1080/10509201003601167. https://doi.org/10.1080/10509201003601167.

Lefkowitz, Melanie. 2019. Study: AI may mask racial disparities in credit, lending. Accessed: 2021-03-31. https://news.cornell.edu/stories/2019/01/study-ai-may-mask-racial-disparities-credit-lending.

Martinuzzi, Alex Webb; Elisa. 2019. The apple card is sexist. blaming the algorithm is proof. Accessed: 2021-03-31. https://www.bloomberg.com/opinion/articles/2019-11-11/is-the-apple-and-goldman-sachs-credit-card-sexist.

McKinsey Institute. 2020. *Global survey: the state of AI in 2020*. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Global%20survey%20The%20state%20of%20AI%20in%202020/Global-survey-The-state-of-AI-in-2020.pdf. Accessed: 2021-03-08.

Nadimpalli, Meenakshi. 2017. Artificial intelligence risks and benefits. *International Journal of Innovative Research in Science, Engineering and Technology* 6(6).

National Geographic. 2018. *City of the future: Singapore – Full episode | National Geographic*. https://www.youtube.com/watch?v=xi6r3hZe5Tg. Accessed: 2021-03-20.

Netflix. 2020. *Netflix quarterly earnings*. https://ir.netflix.net/financials/quarterly-earnings/default.aspx. Accessed: 2021-03-08.

Rawlinson, Kevin. 2019. Digital assistants like Siri and Alexa entrench gender biases, says UN. *The Guardian*. Accessed: 2021-03-25. https://www.theguardian.com/technology/2019/may/22/digital-voice-assistants-siri-alexa-gender-biases-unesco-says.

Skool4Kidz. 2021. https://skool4kidz.com.sg/. Accessed: 2021-03-08.

Statista. 2020a. *Film industry in china - statistics & facts*. https://www.statista.com/topics/5776/film-industry-in-china/. Accessed: 2021-03-20.

Statista. 2020b. *Smart home report 2020*. https://www.statista.com/outlook/dmo/smart-home/worldwide. Accessed: 2021-03-08.

Valentino-DeVries, Jennifer. 2020. How the police use facial recognition, and where it falls short. Accessed: 2021-03-08. https://www.nytimes.com/2020/01/12/technology/facial-recognition-police.html.

Wang, Mei, Weihong Deng, Jiani Hu, Xunqiang Tao & Yaohai Huang. 2019. Racial faces in the wild: reducing racial bias by information maximization adaptation network. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 692–702.

West, Mark, Rebecca Kraut & Han Ei Chew. 2019. *I'd blush if I could: closing gender divides in digital skills through education*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000367416.

Wikipedia contributors. 2020. *List of artificial intelligence films — Wikipedia, the free encyclopedia*. Accessed: 2020-11-22. https://en.wikipedia.org/w/index.php?title=List_of_artificial_intelligence_films&oldid=1009398524.

Chapter 28

AI in literature: An investigation of the science fiction novel QualityLand

Malte Heyen, F.R. & Anita Wagner

How is AI displayed in German literature? We conducted an exemplary analysis of Marc-Uwe Kling's bestseller *QualityLand.* A dual approach, quantitative and qualitative, is used to investigate which AI related concepts are described in the book and how characters in *QualityLand* interact with AI. In the qualitative analysis, ten exemplary concepts are discussed in detail. For the quantitative analysis, the authors of this work extracted all interactions involving AI described in the book. Each interaction was tagged according to the categories *Character*, *Type of AI*, *Type of Interaction*, and *Emotions.* A network analysis was performed based on this data. It is shown that the book *QualityLand* discusses many relevant topics related to AI, reaching from technical aspects over social issues to philosophical debates. Kling describes these concepts intelligible for readers with little to no background knowledge in this area. Additionally, a quantitative analysis of the interactions was performed, specifically for *Peter Arbeitsloser* as the main character of the book. It is found that either more positive or negative sentiments are associated with certain types of interaction and certain AI characters. It can be seen that the book displays natural language processing as a key aspect of interacting with AI, as the main interactions described in *QualityLand* involve conversations. Further, we visualised the connections between the characters in the novel in a network graph.

Keywords: QualityLand | Marc-Uwe Kling | Science fiction | AI in literature | Dystopia | Network analysis

1 Introduction

Literature is still an important medium where public debates take place and current, relevant topics are being discussed. Specifically, popular novels are reaching a wide readership and may set topics or address issues relevant at that time (see Chapter 26 for an analysis of AI in popular movies). Bestsellers might even be discussed outside the author's typical reader demography, or be referenced in other media, forming part of the common culture of a society. Thus, investigating how the topic of AI is displayed in the German literature appears to be highly relevant in the context of the public debate about AI.

Marc-Uwe Kling is a German author, songwriter, and cabaret artist. He became popular through his famous satiric novel *Die Känguru-Chroniken*, published in 2009 as the first part of *Die Känguru Trilogie*. The three books of the trilogy were sold over 2,5 million times as of the beginning of 2018 (Ullstein 2018: p. 153) and the audiobooks were sold over 1,7 million times as of mid 2018 (Börsenblatt 2018b). In 2017, Kling published the science fiction novel *QualityLand*[1] at the 'Ullstein Verlag'. In contrast to his previous work, the story around the protagonist *Peter Arbeitsloser*[2] takes place in the near future in a fictitious country called QualityLand. Kling himself describes it as a 'funny dystopia' (Schillat 2017). In this world, everything is organised by algorithms. Nearly every inhabitant of QualityLand uses a digital personal assistant who knows what they want and decides for them.

The book is available in two versions; a light and a dark one. They only differ in between the chapters, where the reader is presented with recommendations, news and advertisement from QualityLand. They either portray a more positive view (light) or a rather cynical, negative view (dark). *QualityLand* became a Spiegel-Bestseller[3] in 2017 (buchreport 2020) and won the German Science Fiction Award 2018 (Börsenblatt 2018a). In March 2019, HBO announced a series adaptation of *QualityLand* (Goldberg 2019). In 2020, Kling published a second book: *QualityLand 2.0: Kikis Geheimnis*, which again ranked number one Spiegel-Bestseller of the year 2020 in the category audio books (DER SPIEGEL 2021).

Thus, the success of *QualityLand* is undeniable and one reason for us to be curious about how Kling approaches the topic of AI. Another reason is his perspective as a cabaret artist and author, well-known for his comedic dialogues, iconic and ironic word plays, and satiric left-wing social criticism. His readership

[1] In this article, we use *QualityLand* to refer to the book and QualityLand to refer to the land, which is described in the book.

[2] engl.: Peter Unemployed

[3] Ranked 16th in the category 'Hardcover Belletristik' for the year 2017; ranked 1st in the category 'Hörbücher Belletristik/Sachbuch' in edition 40/2017; ranked 4th in the category 'Hardcover Belletristik' in edition 41/2017; ranked 11th in the category 'Taschenbuch Belletristik' in edition 45/2020.

represents a broad demographic population and is presumably not previously educated in the field of AI. As such, *QualityLand* offers a unique perspective on AI, described by the means of a traditional book, combined with an impressive popularity and broad coverage. Altogether, this motivated us to investigate *QualityLand* further and to include it in this work on Artificial Intelligence in public discourse.

2 Methodology

To tackle the problem of working with a novel as our main resource, we decided on a dual approach, combining qualitative and quantitative aspects. We will firstly outline concepts regarding AI and describe how they are being addressed in the book. Secondly, we will investigate how the characters in QualityLand interact with AI. For this, we created a data set including these interactions and performed a network analysis.

2.1 Qualitative analysis: AI concepts

Besides fictional aspects, the book includes and formulates key philosophical and technical concepts related to AI. These are topics which are known in the academic debate about AI, but might not be familiar to the average reader. We will provide examples from the book for ten selected topics to give an impression on how these concepts are presented to the reader. Further, we will give historical, philosophical or technical context for each topic.

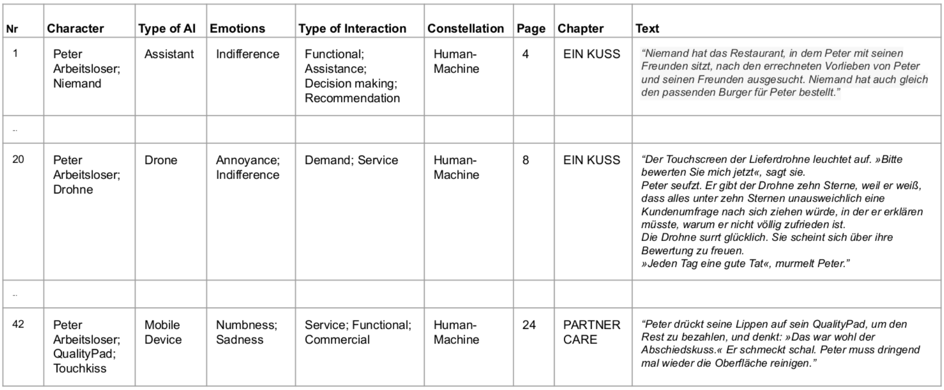
2.2 Quantitative analysis: Data

We generated a data set in which we identified all interactions described in the novel involving AI. Throughout this process, we further identified which types of AI and which types of interaction are described in the book. We also extracted the respective emotions associated to the interactions.[4]

The data generating process was as follows: The three authors of this article read through the book individually and identified text passages describing an interaction involving AI. We added this passage to our data table and (subjectively) tagged the corresponding attributes for each variable. We followed an exploratory approach and iteratively added attributes when deemed appropriate.

[4] The data set can be found at https://docs.google.com/spreadsheets/d/1HtwsYdB9J_M6pYzDt-IFGjrNnD528liX5Gf8F4th7FI/edit#gid=2139048762.

For example, if we encountered a passage where an autonomous drone recommends something to *Peter Arbeitsloser*, we added the attribute *Recommendation* to the variable *Type of Interaction*. The data set consists of four variables: *Character*, *Type of AI*, *Type of Interaction*, and *Emotions*. Each data point consists of these variables of interest, where each variable can have multiple attributes being tagged. Further, we also included the reference to the source, i.e. the chapter, the page, and the text itself. We also added the category *Constellation*, where we tagged whether the interaction was between humans (*Human-Human*), machines including AIs (*Machine-Machine*) or both (*Human-Machine*). But this category was not further used for our analysis [5].



Nr	Character	Type of AI	Emotions	Type of Interaction	Constellation	Page	Chapter	Text
1	Peter Arbeitsloser; Niemand	Assistant	Indifference	Functional; Assistance; Decision making; Recommendation	Human-Machine	4	EIN KUSS	*“Niemand hat das Restaurant, in dem Peter mit seinen Freunden sitzt, nach den errechneten Vorlieben von Peter und seinen Freunden ausgesucht. Niemand hat auch gleich den passenden Burger für Peter bestellt.”*
...								
20	Peter Arbeitsloser; Drohne	Drone	Annoyance; Indifference	Demand; Service	Human-Machine	8	EIN KUSS	*“Der Touchscreen der Lieferdrohne leuchtet auf. »Bitte bewerten Sie mich jetzt«, sagt sie.* *Peter seufzt. Er gibt der Drohne zehn Sterne, weil er weiß, dass alles unter zehn Sternen unausweichlich eine Kundenumfrage nach sich ziehen würde, in der er erklären müsste, warum er nicht völlig zufrieden ist.* *Die Drohne surrt glücklich. Sie scheint sich über ihre Bewertung zu freuen.* *»Jeden Tag eine gute Tat«, murmelt Peter.”*
...								
42	Peter Arbeitsloser; QualityPad; Touchkiss	Mobile Device	Numbness; Sadness	Service; Functional; Commercial	Human-Machine	24	PARTNER CARE	*“Peter drückt seine Lippen auf sein QualityPad, um den Rest zu bezahlen, und denkt: »Das war wohl der Abschiedskuss.« Er schmeckt schal. Peter muss dringend mal wieder die Oberfläche reinigen.”*

Figure 1: Example of the data set

In the following we will describe the five categories of our tag set. In the category *Character* we identified the individuals involved in the situation. They include humans like the protagonist *Peter Arbeitsloser*, androids that appear as main characters such as *John of Us*, other robots like the *Mülleimer*[6] and mere objects (e.g. *Intelligente Tür*[7]) who also interact and have a personality. If an AI was involved in the situation, we added the appropriate attribute in the category *Type of AI*. For example, when *John of Us* appeared we added the attribute *android*. Further, we ascribed emotions like *Annoyance* and *Joy* to the situation. We did not differentiate whether a single character had a certain emotion or

[5] Because the large majority of interactions in the book are *Human-Machine*, we primarily focused on those and only partially included the others. Thus, our data set consist of all interactions with AI, and some interactions which display AI in general, but do not involve characters directly interacting with an AI.

[6] engl.: rubbish bin

[7] engl.: smart door

whether it was an emotion ‘present’ in the interaction. The attribute *NONE* was used when no specific emotion could be inferred. To specifically describe the interaction, we tagged attributes to the category *Type of Interaction*. For instance this could be the attribute *Informative*, when one character provides information to another character. Other interactions include *Social* or *Conversation*, which would be characters dining at a restaurant or having a longer dialogue.

Thus, we end up with three data sets, one for each contributing author, which each cover the whole book. The three data sets have 100, 151 and 179 entries. Combining them results in a data set of length 430. The merging should not overrepresent single attributes over others, but reduce the weighting of potential outliers and human error. However, some interactions might only be tagged by some authors, which would underrepresent them in the data set. These interactions might also be less clear identifiable and results in a weighting for certainty. The same holds for the attributes, as attributes which are very clearly present in an interaction should be tagged by all three authors, while attributes which are less certain might only be identified by one or two authors.

2.3 Quantitative analysis: Interactions with AI

We decided to use a method for our data analysis which is based on a social network analysis. These types of analyses are usually used to investigate how agents in a network interact or are related to each other. This network can be displayed as a graph, containing nodes as agents and edges as connections between agents. In our analysis, the attributes are handled similar to agents. Each attribute appears together with other attributes, and the connection between two attributes is based on the number of common occurrences. In our graphs, each node is a representation of an attribute and the number of common occurrences with other attributes is the weight of the connecting edge between the nodes. An example for a resulting graph can be seen in Figure 5. To investigate the interactions further, we extracted cliques out of our graph. A clique is a subset of nodes where each node has a connection to all the other nodes in this clique. We calculated the maximal cliques. These are constructed in such a way that there is no node outside of the clique, which has a connection to all the nodes inside of the clique. Thus, it cannot be extended by including more nodes. As such, it is a clique which cannot be merely a subset of another, larger clique. Each clique is a fully connected subgraph. We calculated a total edge weight for each clique as the sum over all the weights of the edges. From this, we derived the mean edge weight for each clique. This allowed us to investigate, how strong the average connection between attributes is within a clique. The higher the mean edge weight, the higher the average frequency of common appearances of the attributes. We encountered the issue, that we could not calculate optimal cliques (which maximise the mean edge weight). Calculating optimal cliques might have led to a better understanding of the connection between attributes. However, calculating optimal cliques lies within the complexity class of NP-hard, and we therefore decided to use the maximal cliques instead.[8]

3 Qualitative results: Ten examples of AI concepts

In the following, we will present key philosophical and technical concepts related to AI and how they are displayed in *QualityLand*. We provide ten examples from the book and try to give historical, philosophical or technical context for each topic. The examples are presented as in the original source in German, but translations are available as footnotes.

3.1 Turing Test

In 1950, Alan Turing published his famous article on 'Computing machinery and intelligence', in which he raises the question: Can machines think? He presents his idea on how to approach this question with his 'imitation game', today better known as Turing test (Turing 1950). In *QualityLand*, this idea is presented through the character *'der Alte'* (engl.: the old [man]). He not only explains the Turing test to the character *Peter Arbeitsloser*, he also takes the thoughts further with the claim, that "an AI that passes the Turing Test would also be intelligent enough to fail it on purpose." (Kling 2017: p. 185). He describes the procedure as following:

> »Was ist der Turing-Test?«
>
> »Alan Turing hat 1950 eine Methode vorgeschlagen, mit der man angeblich feststellen kann, ob eine Maschine ein dem Menschen gleichwertiges Denkvermögen hat.«
>
> »Und wie soll das gehen?«
>
> »Ein Mensch bekommt zwei Gesprächspartner, die er weder hören noch sehen kann. Kommuniziert wird per Tastatur. Der eine Gesprächspartner ist ein Mensch, der andere eine künstliche Intelligenz. Gelänge es dem Fragesteller nicht, herauszufinden, welcher seiner Gesprächspartner Mensch und

[8] The code for our analysis can be found at https://gitlab.com/FRadtke/aipd_qualityland_analysis/-/blob/master/AIPD_Qualityland_Analysis.ipynb.

> welcher Maschine ist, dann hätte die K. I. ein dem Menschen ebenbürtiges Denkvermögen.« (Kling 2017: p. 185)[9]

Later on, *'der Alte'* explains 'Captcha' (Completely Automated Public Turing Test to Tell Computers and Humans apart) which refer to image recognition tests (Kling 2017: p. 229). A typical example of such a test is the correct identification of contorted letters or certain objects in a set of images by either a human or a machine. As Kling portrays it in the book, image recognition improved so significantly in the near future and machines got better than humans at these tests. Thus, humans and machines can still be distinguished, only that the one passing the test now is the machine.

3.2 Weak & Strong AI

The scientific (and science fiction) literature often distinguishes between the terms *Weak AI* and *Strong AI*. Commonly they refer to narrow, task specific AI systems (weak) and those which have general problem solving capacities (strong). This is also how they are introduced in *QualityLand*, in a dialogue between *Peter* and *'der Alte'*:

> »Ist dir der Unterschied zwischen einer schwachen und einer starken künstlichen Intelligenz geläufig?«
>
> »Ganz grob«, sagt Peter. »Eine schwache K. I. ist für eine spezifische Aufgabe konstruiert. Zum Beispiel, ein Auto zu lenken. Oder für die Rücknahme unerwünschter Produkte. Und sie kann sehr nervig sein.«
>
> »Ja, so ungefähr. Und eine starke K. I.?«
>
> »Eine starke künstliche Intelligenz wäre eine K. I., die nicht speziell für eine Aufgabe programmiert werden muss. Eine allgemeine Problemlösungsmaschine, die erfolgreich jede intellektuelle Aufgabe ausführen kann, die ein Mensch meistern könnte. Die vielleicht sogar ein echtes Bewusstsein hätte. Aber so etwas gibt es nicht.« (Kling 2017: p. 184)[10]

[9] engl.: "What is the Turing test?" "In 1950, Alan Turing suggested a method to determine whether a machine has the same intellecutal power as humans" "And how does it work?" "A human being has two conversational partners, that he can neither see nor hear. They communicate via a keyboard. One of the conversational partners is a human and the other one is an AI. If the questioner is not able to tell human and machine apart, then the AI could be considered to have the same intellectual power as humans."

[10] engl.: "Are you aware of the difference between a weak and a strong AI?" "Very roughly" says Peter. "A weak AI is built for a specific task. For example a car to navigate. Or for the retraction

However, the term *Strong AI* in the philosophical debate, as originally coined by Searle in 1980, does not focus on the AI's capacities at various tasks, e.g. playing chess (Searle 1980). Rather, according to Searle, *Strong AI* is a philosophical position which holds that the mind is essentially a computer program and the brain is a (biological) computer which can be emulated. It also follows that computer systems can have mental states and possess consciousness, which has been famously opposed by Searle's Chinese Room experiment (see Searle 1980). *Weak AI*, in contrast, is a position which only claims that an AI system can exhibit intelligent behaviour, without assuming mental states or consciousness. We can observe that the scientific definitions diverge from the presentations in the book (Kling 2017: pp. 184-191). *'der Alte'* defines strong AI as "an AI which can do anything a human can do. But faster, of course. Without errors." (Kling 2017: pp. 185-186)[11]

3.3 Superintelligence

Nick Bostrom defines superintelligence (SI) as "any intellect that is[sic] vastly outperforms the best human brains in practically every field, including scientific creativity, general wisdom, and social skills" (Bostrom 2003: p.1). SI is often thought of as emerging from an intelligence explosion, as originally described by I. J. Good in 1965 (Good 1965). This is also how we encounter the concept of SI in *QualityLand*. In the same dialogue between *Peter* and *'der Alte'* as referenced above, *'der Alte'* describes the process as following:

> Eine starke K. I. ist eine intelligente Maschine, der es möglich ist, eine noch intelligentere Maschine zu entwerfen, welcher es wiederum möglich sein wird, eine noch viel intelligentere Maschine zu entwerfen. Rekursive Selbstverbesserung. Es käme zu einer Intelligenzexplosion! [...] Eine Superintelligenz entstünde. (Kling 2017: p. 186)[12]

In this context, Kling also mentions a religious aspect in the debate about SI by comparing the SI to god:

of unwanted products. And she can be very annoying." "Yes, almost. And what about strong AI?" "A strong AI would be an AI that does not need to be programmed for a specific task. It is a machine to solve general problems, that is able to execute every intellectual task that a human is able to cope with. Maybe it would even have a real consciousness. But something like that does not exist"

[11] ger.: Eine K. I., die alles kann, was ein Mensch kann. Nur schneller natürlich. Ohne Fehler.

[12] engl: A strong AI is able to build a machine which is more intelligent than itself, and this one can build an even more intelligent one, and so forth. This would be a recursive self-enhancement leading to an explosion of intelligence, a so called SI.

> »[…] Nun sag mir, wie nennst du ein Wesen, das allgegenwärtig, allwissend und allmächtig ist?«
>
> »Gott?«, fragt Peter. (Kling 2017: p. 187)[13]

3.4 Moravec's paradox

An observation in AI and robotics research is, as Hans Peter Moravec describes it in 1988, that "it is comparatively easy to make computers exhibit adult-level performance in solving problems on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility" (Moravec 1995: p. 15). This observation is called *Moravec's paradox* and is referenced in *QualityLand* multiple times, inter alia in the appropriately titled chapter *'Das Moravec'sche Paradox'*. Here, the highly advanced android *John of Us* is struggling to simply carry a cup of coffee.

> »John übt, eine volle Kaffeetasse quer durch den Raum zu tragen«, sagt der Trainer. »Wir machen beachtliche Fortschritte!«
>
> Die Frau wendet sich an Tony.»Sie wollen die Staatsgeschäfte jemandem übertragen, der nicht mal eine Tasse tragen kann, ohne dabei zu kleckern?«
>
> John sieht sie scharf an.
>
> »Man nennt es das Moravec'sche Paradox«, sagt er.
>
> »Tut man das?«
>
> »Hans Moravec war ein Pionier im Bereich der künstlichen Intelligenz«, sagt John. »Er musste feststellen, dass für eine K. I. die schwierigen Probleme einfach und die einfachen Probleme schwierig sind. Scheinbar simpelste Aufgaben, wie die sensomotorischen Fähigkeiten eines Einjährigen oder, nun ja, das Tragen einer vollen Tasse, kosten eine K. I. unglaublich viel Rechenleistung, während scheinbar komplizierte Aufgaben, wie das Schlagen eines Schachgroßmeisters, für eine K. I. recht simpel sind.« (Kling 2017: p. 75)[14]

[13] engl.: "Tell me, how do you call a being that is omnipresent, omniscient and omnipotent?" "God?" asks Peter.

[14] engl.: "John practises to transport a full cup of coffee across the room", says the trainer. "We are making considerable progress!" The woman turns to Tony. "You want to assign state affairs to someone, who is not even capable of carrying a cup without spilling?" John gives her a sharp look. "You call it the Moravec's paradox", he says. "You do?" "Hans Moravec was a pioneer in the field of AI", says John. "He needed to realise, that for an AI hard problems are simple and simple problems are hard. Seemingly simplest tasks, like the sensorimotor skills of a one year old or the carrying of a full cup, cost an AI full computing power, whereas seemingly complex tasks, such as defeating a chess grand master, are quite simple for an AI."

3.5 Uncanny valley for androids

The concept was first identified by the robotics professor Masahiro Mori as *bukimi no tani genshō* in 1970, which was translated as uncanny valley (Mori et al. 2012). His original hypothesis states that if the appearance of a robot is made more human, humans' emotional response to the robot becomes more positive and empathetic. However, at a certain point, the response starts to become revulsion. Further changing the robot's appearance to become even more human-like, normalizes the emotional response and it becomes positive again until it approaches human-to-human level. The concept of an uncanny valley is not explicitly mentioned in the book, but it is apparent in some interactions between humans and androids:

> Erst auf den zweiten Blick erkennt Sandra, dass es sich um eine Maschine handelt.
>
> »Täuschend echt, diese neuen Androiden, was?«, fragt Peter.
>
> »Ja. Fast ein bisschen unheimlich«, sagt Sandra. (Kling 2017: p. 33)[15]

3.6 Replacement of jobs

For several years, there is a huge debate to which extent human jobs will be replaced by AI. Some AI experts estimate that within the next 10 - 15 years between 40% - 50% of the current jobs executed by humans will be transferred to AIs (Maier 2017, Reisinger 2019). In the book, this consequence of AI applications is especially addressed in a scene, where *'der Alte'* guides *Peter* through a thinking process of the potential implications of these replacements. In the end the question is what would happen if an AI could do everything better than humans. The answer given by *Peter* is that all his problems and even his existence would be meaningless (Kling 2017: p.188). This shows the social and psychological impact AI may have. A concrete example for this development in the book are autonomous cars which almost completely replace human drivers.

3.7 Autonomous cars and the trolley problem

Self-driving cars are a recurrent theme in *QualityLand* and they, besides driving, often interact with their passengers. Most citizens in QualityLand do not own a car, but simply order one via their QualityPad. Peter Arbeitsloser interacts with

[15] engl.: Only at second glance Sandra realizes that it is a machine. "These new androids are deceptively real, aren't they?", asks Peter."Yes. Almost a bit scary", says Sandra.

three different cars. They are able to display different personality traits, e.g. choleric or talkative. The car *Carl* holds small talk with *Peter* while driving and both exchange music interests. *David* is another self-driving car; a ghost car, who lost his sense of orientation due to a malfunctioning navigation system. He is not able to navigate on his own and requires guidance by his passengers. The cars are described in a very anthropomorphic manner, including error-proneness, such as losing the sense of orientation, and showing human personality traits.

Further, the cars show an interest in philosophical debates. The classical trolley problem is specifically addressed in the chapter *"Moralische Implikationen"*[16] by the autonomous car *Herbert* who discusses this issue with *Peter*. The trolley problem (Foot 1967) describes a moral dilemma in which a person has to decide to act (e.g. changing a steering switch) in order to rescue a group of individuals by sacrificing another (often smaller) group. It has often been revisited in the context of autonomous cars (Lim & Taeihagh 2019). Say, an autonomous car is in a situation in which it can only rescue the driver by driving into a group of pedestrians. The car should be programmed beforehand to behave morally in this scenario. But what this means and how it can be implemented is still subject to ethical and scientific debate. These considerations are part of a conversation between *Peter* and *Herbert*, in which they talk about car accidents:

> »Also nicht, dass ich je einen gebaut hätte«, sagt das Auto lachend. »Es sind mehr die moralischen Implikationen eines Unfalls, die mich faszinieren.«
>
> »Wie meinst du das?«
>
> »Nun«, sagt Herbert, »für einen Menschen ist ein Unfall nur sehr selten mit einer moralischen Entscheidung verknüpft. Eure Denkprozesse sind zu langsam. [...] Eine Maschine allerdings reagiert viel schneller und hat Zeit für genau diese komplexen Überlegungen. Für uns beinhaltet fast jeder Unfall eine moralische Entscheidung.« (Kling 2017: pp. 154-155) [17]

3.8 Surveillance

Most of the public areas in QualityCity, the capital of QualityLand, are monitored by security cameras (Kling 2017: p. 249). Throughout the whole book, the issue

[16] engl.: moral implications

[17] "Not that I ever caused one", laughs the car. "It is more the moral implications of an accident that fascinate me." "How do you mean?" "Well", says Herbert, "for humans an accident is quite rarely combined with a moral decision. Your thought processes are too slow. [...] A machine reacts way faster and has the capacity for this complex considerations. For us almost every accident contains a moral decision."

of surveillance and privacy is addressed repeatedly. One example is the character *Kiki Unbekannt*, who tries to stay under the radar and behaves as unpredictable for the algorithms as possible. In the chapter *'Zack'*, she hacks an autonomous car and uses a hijacked profile to travel anonymously (Kling 2017: p. 162). But even in their private homes, companies and the government collect data from the inhabitants of QualityLand. *Peter* often stands in his scrap press, where all communication signals are blocked, presumably to have a moment of privacy. Further, potential misuse and problems with data security are mentioned. After being hacked and an explicit video of him was published, *Martyn Vorstand* tries to retake his privacy by covering all the cameras in his house (Kling 2017: p. 249). Video surveillance in public and private space is a often discussed topic in public media. Especially since systems like SkyNet in China (Zhang 2019) or the CCTV system in London have been implemented. A social scoring system, similar to SkyNet, is described in *QualityLand*. Citizen are assigned a *'level'*. The society is divided and ranked according to this level. A higher level comes with certain privileges, whereas people below level ten belong to the so-called 'useless'.

3.9 Personalization and filter bubbles

In QualityLand, algorithms are responsible for almost all important (life) decisions of the citizens and filter most of the information they receive. They are especially used for recommendations and advertisement. Not only products are recommended, but also friends and partners. *Peter*'s friends, for example, were recommended to him by *Niemand*. Personal advertisement is delivered through Apps on his QualityPad. The algorithm from *TheShop* is supposed to know what the customer wants and directly orders the product. However, in *Peter*'s case, this went wrong and he was delivered a pink vibrator in the form of a dolphin. During the whole story, *Peter* attempts to return the pink vibrator, and tries to prove the system wrong, represented by *TheShop* and other companies. However, it turns out to be nearly impossible to influence the algorithms. *Peter* is often annoyed that he does not receive new input, like new music or relevant political information. This is because the algorithms are biased by his former behaviour and his profile as a 'useless'. This shows how filter-bubbles manifest themselves like a self-full-filling prophecy, and how difficult it is to escape them. To read more on personalization algorithms see Chapter 12.

3.10 Bias in algorithms and discrimination

Discrimination through automated systems has become an increasingly social issue in recent years. A prominent example was Microsoft's chatbot Tay, which was released in March 2016 (Schwartz 2019). Within a couple of hours, Tay adopted hate-speech and produced highly offensive tweets. The reason for this was the data on which the AI was trained. It also contained racist and antisemitic language, coming from users and other bots. While this example is still relatively harmless, it becomes more serious if we think about other domains where important automated decisions are being made by potentially discriminating AI systems. In *QualityLand*, this issue is repeatedly addressed. One example can be found in a dialogue between *Peter Arbeitsloser* and *'der Alte'*, where they discuss the application of AI systems for selecting job applicants:

> »Ach, selbst wenn alle Profile stimmten, würden uns die Algorithmen noch diskriminieren.«
>
> »Aber warum?«, fragt Peter. »Müssten Maschinen nicht objektiv sein?«
>
> »Unsinn«, sagt der Alte. »Folgendes Beispiel: Ein Human-Resources-Algorithmus lernt, indem er die zahlreichen Entscheidungen durchforstet, die menschliche Personalmanager vor ihm getroffen haben. Er stellt fest, dass Bewerber mit schwarzer Hautfarbe überproportional selten eingestellt wurden. Es ist also nur logisch, Bewerber mit schwarzer Hautfarbe gar nicht erst einzuladen. Verstehst du? Wenn man vorne in einen Algorithmus Vorurteile hineinsteckt, kommen hinten Vorurteile heraus.« (Kling 2017: p. 206) [18]

Another example for discrimination can be found when the autonomous car *Carl* refuses to drive Peter to his home, because the area where he is living is assigned as a high risk territory for the car. The economic interest of the company is the reason why *Peter* is being discriminated based on the area where he lives. Ironically, robots and androids are also being discriminated in QualityLand by parts of the human society (see e.g. Kling 2017: p. 57,p. 279). As AI characters in the book are displayed very human-like, they are easy to relate and sympathise with. This contributes to emphasising the issue of discrimination in society.

[18] "Even if all profiles were correct, algorithms would still discriminate us." "But why?", asks Peter. "Wouldn't machines have to be objective?" "That's nonsense", says der Alte. "Here is an example: a human-resources-algorithm learns while trawling through numerous decisions that human resources managers made before him. He finds that applicants with black skin color were hired disproportionately infrequently. From this, it's only logical not to invite black applicants at all. You get it? If you insert prejudices in an algorithm prejudices will emerge from the algorithm."

4 Quantitative results: Interactions with AI

In the following section, we will investigate the generated data set further. Firstly, we will have a look at the occurrences of attributes for the different categories. Secondly, we examine the network graph visualising the connections between characters. Afterwards, we try to approach an interpretation of the calculated cliques.

4.1 Occurrences of attributes

Peter Arbeitsloser is the main character of the book. As such, he is also the prevalent character found in the interactions (see Figure 2). The most often appearing characters in the interactions are further: *Kalliope 7.3* (N=79), *John of Us* (N=56), *Niemand* (N=54), and *Pink* (N=48).

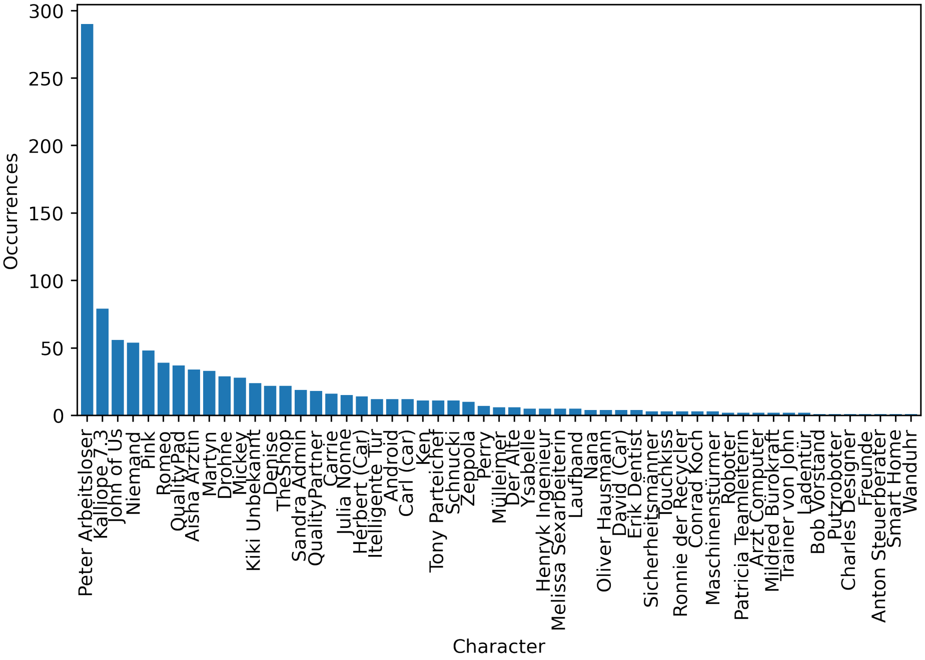


Figure 2: Occurrences of characters

In the following, we will have a closer look into the interactions of *Peter* (N=291). The five AI characters he interacts with most often are *Kalliope 7.3* (N=71), *Niemand* (N=45), *Pink* (N=38), *Romeo* (N=35), and *QualityPads* (N=29). In the 45 interactions with *Niemand*, the four most tagged emotions are *NONE* (N=12), *Annoyance* (N=7), *Anger* (N=7), and *Numbness* (N=4). In the 29 interactions with QualityPads, we find the four most tagged emotions to be *Annoyance* (N=10), *Sadness* (N=4), *Shame* (N=4), and *Numbness* (N=4). From this, we can argue that he has a rather neutral or negative sentiment towards his win-assistant *Niemand* and QualityPads. In contrast, we can see that *Peter* has a rather positive relationship with *Kalliope 7.3*, *Pink*, and *Romeo*. In the 71 interactions with *Kalliope 7.3*, we find the four most tagged emotions to be *Amusement* (N=15), *Joy* (N=13), *Trust* (N=13), and *NONE* (12). In the 38 interactions with *Pink*, the four most tagged emotions are *Amusement* (N=13), *Excitement* (N=10), *Trust* (N=10), and *Happiness* (N=9). The four most tagged emotions in the 35 interactions with *Romeo* are *Amusement* (N=12), *Excitement* (N=10), *Happiness* (N=9), and *Interest* (N=9).

It is important to note again, that these emotions are not directed (e.g. from *Peter* to *Kalliope 7.3*) but only identified to be present in the interaction. We conclude that depending on the AI character the interactions that Peter has are either associated with positive or negative emotions. The interactions with his AI friends (represented by *Kalliope 7.3*, *Pink*, and *Romeo*) are displayed very differently to the interactions with more functional AI assistants (e.g. *Niemand* and *QualityPads*). This difference in associated emotions can also be found if we look at the different types of interactions.

The type of interaction which occurs most often is *Conversation*, followed by *Social* and *Supportive*. Only after these occur *Informative* and *Functional*. In line with this, while reading we noticed that natural language processing seems to be a key feature of most AIs described in the book. One explanation for this may be that AI characters are displayed very human like and are present in many social interactions. These often include conversations. Another reason might be Kling's writing style, which is known for including many dialogues in general.

We can now compare the prevalent types of interactions based on the associated emotions. The four emotions most often associated to interactions tagged as *Conversation* are *Anger* (N=29), *Surprise* (N=25), *Amusement* (N=23), and *Annoyance* (N=19). Looking only at the interactions of the type *Social*, the emotions most often associated are *Surprise* (N=27), *Amusement* (N=21), *Happiness* (N=20), and *Annoyance* (N=20). This is very similar to the emotions found for *Conversation*. However, it seems that many interactions which are conversations, but that do not have a social aspect are associated with the emotion *Anger*.

The emotions most often associated with *Supportive* interactions are *NONE* (N=16), *Surprise* (N=15), *Amusement* (N=14), and *Trust* (N=14). In comparison, looking at *Functional* interactions, the four emotions most often associated with these are *NONE* (N=12), *Anger* (N=8), *Indifference* (N=8), and *Annoyance* (N=7).

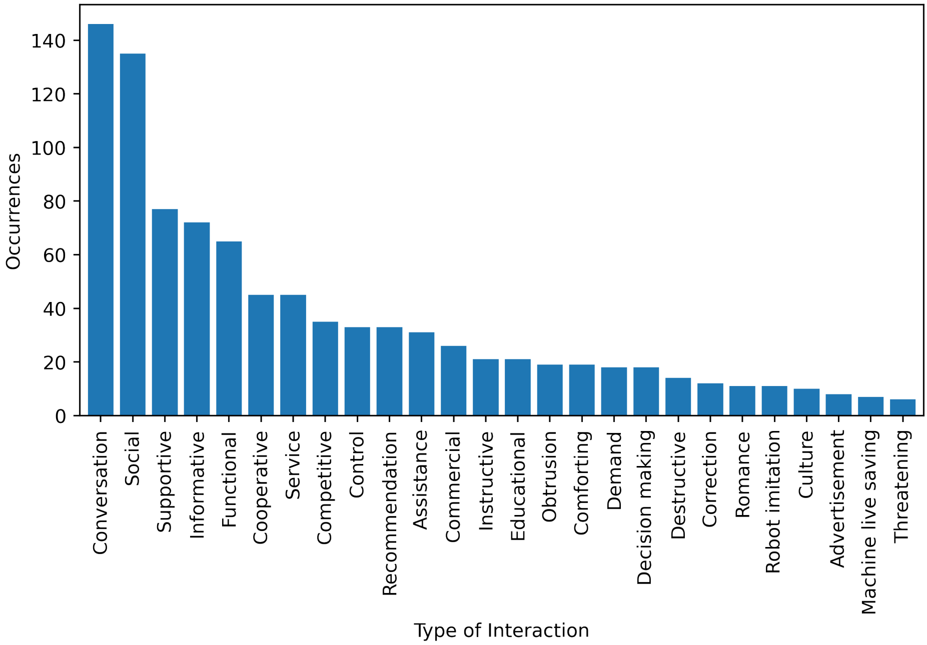


Figure 3: Occurrences of attributes for *Type of Interaction*

For *Informative* interactions, these are quite similar, namely *Annoyance* (N=16), *NONE* (N=12), *Surprise* (N=11), and *Anger* (N=9). We can conclude that these different types of interactions are associated with either positive emotions or negative emotions.

The types of AI which occur most often are *Android*, *Assistant*, *Robot*, and *Algorithm* (see Figure 4). This is not surprising, as both most frequently occurring characters after *Peter*, *Kalliope 7.3* (N=79) and *John of Us*, are androids.

4.2 Network graph

The network graph in Figure 5 visualises the relations of the different characters in the book. We can identify one main cluster and two smaller ones. First, one can see that *Peter Arbeitsloser* as the main character is at the centre of the network and his node is connected to most other characters. Further, we can observe the importance of *John of Us* and *Martyn*, who have separate, smaller clusters. This is because both characters are also very prevalent in the book, and are involved in most interactions. Several chapters are written from their perspectives and

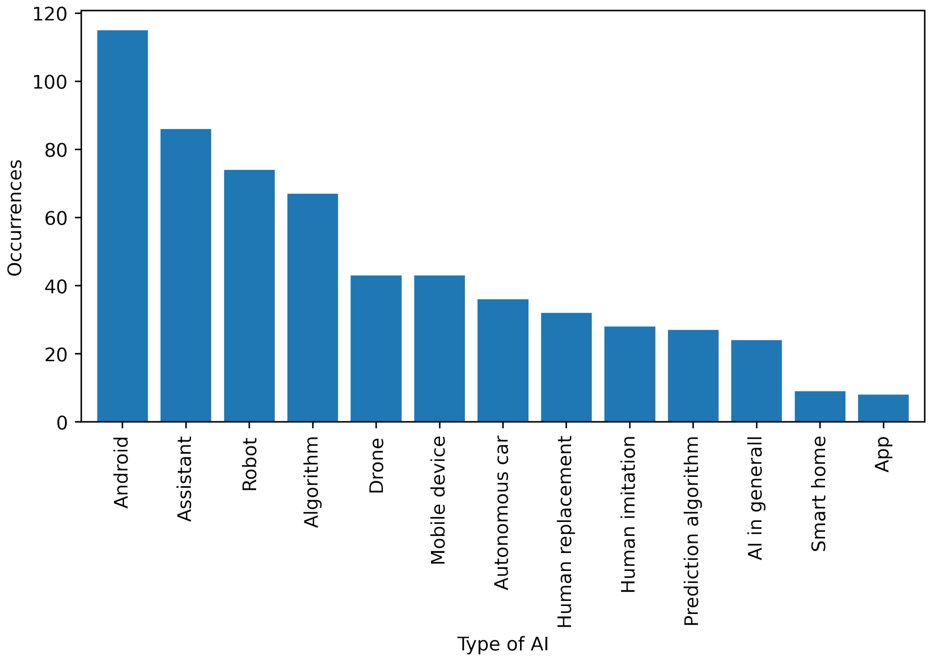


Figure 4: Occurrences of attributes for *Type of AI*

they only meet *Peter* once in one of the last chapters. Thus, they generated two separate clusters.

4.3 Cliques of attributes

We will exemplary present the cliques regarding *Peter Arbeitsloser*. In the clique with the category combination *Character* and *Type of Interaction* we have the maximal clique: [*Freunde*, *Peter Arbeitsloser*, *Social*]. In the category combination *Character, Type of Interaction* and *Type of AI* we have the same attributes plus the attribute *algorithm*. The combination of the categories *Character* , *Type of Interaction* and *Emotions* yields the clique [*Peter Arbeitsloser*, *Ronnie der Recycler*, *Kalliope 7.3*, *Pink*, *Romeo*, *Conversation*, *Social*, *Shame*]. If we add the category *Type of AI* the clique is increased by the attributes *Robot* and *Human imitation*.

We calculated several different cliques involving all combinations of the categories for different attributes e.g. *Autonomous car* or *Android*. However, the resulting cliques do not seem to give further insights into the interactions and do not provide any directly interpretable results. We thus excluded the further analysis of the cliques from this publication.

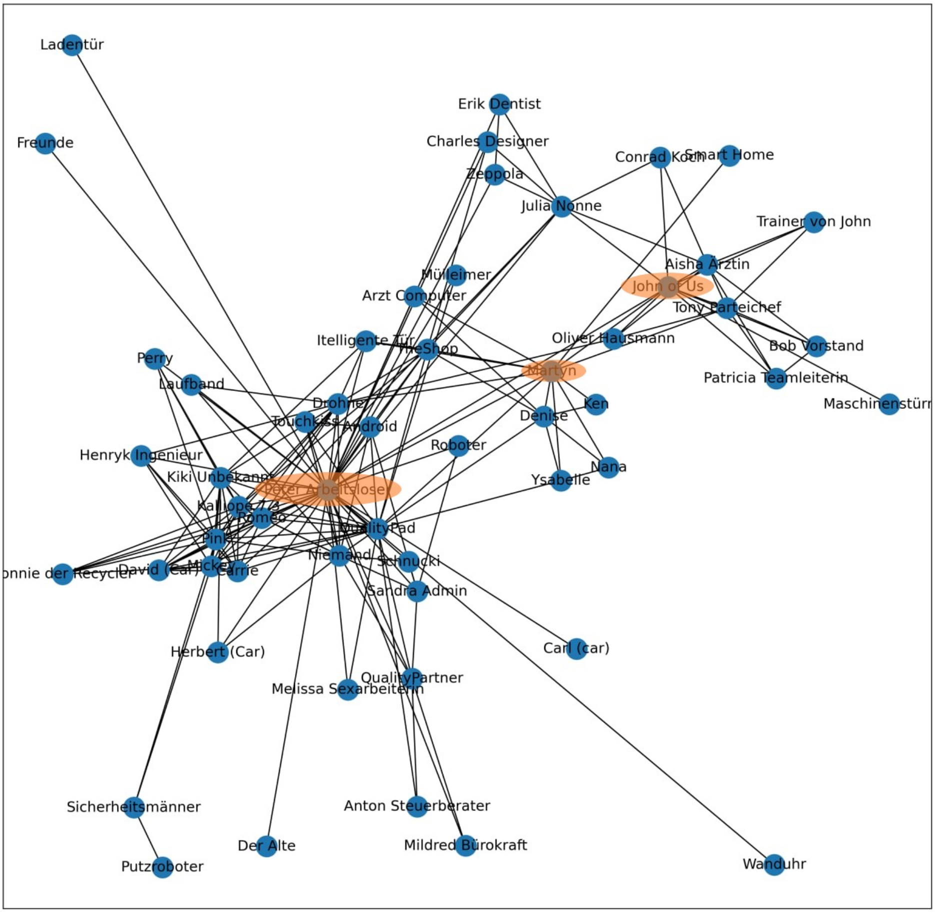


Figure 5: Network including all characters involved in interactions with AI.

5 Discussion

In the following we will shortly portray limitations of our work and conclude with a short summary of our findings.

5.1 Limitations

Regarding our approach and methodology, we see several limitations which shall be addressed. In the qualitative approach our results are not exhaustive, but only portray a selection of relevant topics. Defining inclusion criteria and a more structured approach could improve comparability with other literature and replicability.

Regarding the data generating process, we find that the exploratory approach for adding attributes comes with certain restrictions. We could have defined stricter definitions for the attributes of the different categories, i.e. to have a clear distinction between *Functional* and *Supportive*. Another aspect that could be improved in this context is the huge variance between the raters. We did not compare our results in between, however an inter-rater agreement after tagging the first pages would very likely have led to more comparable results. Calculating a inter-rater reliability would have allowed us to quantify how severe the divergence between raters actually is. An example of this can be found in Chapter 6.

For the categories *Emotion* and *Type of Interaction* we acknowledge the lack of adding a direction to the attribute. For example emotions like *Hate* and *Jealousy* could be ascribed to a certain character who has this emotion and potentially also to the counterpart of the interaction whom the emotion is directed towards. Likewise a direction could have been added to types of interaction such as *Supportive* and *Demanding*.

Regarding the result of our network analysis, we realized that the calculated cliques were not directly interpretable. Attributes which occurred only once in our data set were present in many of the cliques and we inferred that these are not representative findings. Hence, we conclude that the cliques were not suitable for our analysis and we were not able to execute a meaningful interpretation.

5.2 Conclusion

We conclude that many relevant topics related to AI are addressed in the book, reaching from technical aspects over social issues to philosophical debates. Kling describes these concepts intelligible for readers with little to no background knowledge in AI. We were surprised that so many relevant aspects of AI are included in a quite precise manner. However, probably due to the narrative nature of being a novel, Kling does not necessarily hold on to scientific definitions. A key rhetoric instrument he uses is anthropomorphism. Thus, the AI characters in *QualityLand* are being displayed as having intentions, emotions and social relationships. This might lead to a misconception among the readers between what robots and AI systems are currently capable of and the fictitious and ironic features in the book. Nevertheless, Kling might evoke interest in AI for some of his readers.

Regarding the quantitative analysis, we find that the interactions with AI characters are displayed quite versatile. Surprisingly many interactions include conversations and natural language processing seems to be a key feature of most AIs described in the book. We find that either more positive or negative sentiments are associated with different types of interactions and different AI characters. Based on the example of *Peter Arbeitsloser*, we conclude that this is rather an aspect of him having different relationships with different characters. These relationships are then associated with certain types of interactions (e.g. social or functional) and certain emotions (e.g. trust or annoyance). It could further be investigated if this association pattern holds across characters of the same type of AI or is more strongly correlated to the type of interaction. However, it was generally quite difficult to interpret the results of our quantitative analysis. Clear observations, like identifying main characters and main types of interactions, could be reported but further insights can not really be drawn from the quantitative methods we applied.

The same holds for the network analysis and the resulting graphs. We were able to visualise the connections between the characters but could not draw any further insights from this regarding how AI is displayed in the novel. A further analysis might involve grouping the characters and calculating the edge weights not based on frequency but based on the associated emotions or type of interaction.

To get a broader view on the topic 'AI in literature' further analysis of other books involving AI could be conducted. This would probably need the addition of other methods. The generated data set in this project could also be used for other analysis. However, these are out of the scope for this publication.

References

Börsenblatt. 2018a. *"QualityLand" ist der beste SF-Roman.* https://www.boersenblatt.net/archiv/1500477.html. Accessed: 2021-03-15.

Börsenblatt. 2018b. *Känguru wieder im Goldrausch.* https://www.boersenblatt.net/archiv/1544582.html. Accessed: 2021-03-24.

Bostrom, Nick. 2003. Ethical issues in advanced artificial intelligence. *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence* 2. 12–17.

buchreport. 2020. *Jahresbestseller 2020 - buchreport.* https://www.buchreport.de/spiegel-bestseller/jahresbestseller/hoerbuch/. Accessed: 2021-03-12.

DER SPIEGEL. 2021. *Bestseller: Hörbuch Belletristik.* https://www.spiegel.de/kultur/literatur/spiegel-bestseller-hoerbuecher-a-1025528.html. Accessed: 2021-03-11.

Foot, Philippa. 1967. The problem of abortion and the doctrine of the double effect. *Oxford Review* 5(5). 5–15.

Goldberg, Lesley. 2019. *Mike Judge solidifies HBO future with rich overall deal, pair of series orders.* https://www.hollywoodreporter.com/live-feed/mike-judge-solidifies-hbo-future-rich-deal-pair-series-orders-1193760. Accessed: 2021-03-12.

Good, Irving John. 1965. Speculations concerning the first ultraintelligent machine. *Advances in computers* 6(99). 31–83.

Kling, Marc Uwe. 2017. *Qualityland (dark version).* Berlin: Ullstein Tb.

Lim, Hazel Si Min & Araz Taeihagh. 2019. Algorithmic decision-making in AVs: Understanding ethical and technical concerns for smart cities. *Sustainability* 11(20). https://www.mdpi.com/2071-1050/11/20/5791.

Maier, Florian. 2017. *Werden Sie durch KI ersetzt?* https://www.computerwoche.de/a/werden-sie-durch-ki-ersetzt,3331978. Accessed: 2021-03-26.

Moravec, Hans. 1995. *Mind children: The future of robot and human intelligence.* 4. print. Cambridge: Harvard Univ. Press.

Mori, Masahiro, Karl F MacDorman & Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19(2). 98–100.

Reisinger, Don. 2019. *A.I. expert says automation could replace 40% of jobs in 15 years.* https://fortune.com/2019/01/10/automation-replace-jobs/. Accessed: 2021-03-04.

Schillat, Florian. 2017. *Bestseller "QualityLand": Die Zukunft war früher auch besser.* https://www.stern.de/neon/feierabend/-qualityland--von-marc-uwe-kling--die-zukunft-war-frueher-auch-besser-7665760.html. Accessed: 2021-03-07.

Schwartz, Oscar. 2019. *In 2016, Microsoft's racist chatbot revealed the dangers of online vonversation.* https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation. Accessed: 2021-03-04.

Searle, John Rogers. 1980. Minds, brains, and programs. *Behavioral and brain sciences* 3(3). 417–424.

Turing, Alan M. 1950. Computing machinery and intelligence. *Mind* 59(236). 433.

Ullstein. 2018. *Ullstein Taschenbuch Herbst 2018*. Ullstein Buchverlage GmbH. https://issuu.com/ullsteinbuchverlage/docs/utb_h18_gesamtlayout_web_es_9dec24f01eb70c.

Zhang, Phoebe. 2019. *Cities in China most monitored in the world, report finds*. https://www.scmp.com/news/china/society/article/3023455/report-finds-cities-china-most-monitored-world. Accessed: 2021-03-04.

Artificial Intelligence in Public Discourse

Student editors: J. Näther, M. Richter, M. Ruland, A. Wiedenroth

Part I: The general public

Part II: AI and subject experts

Part III: The media

Part IV: Politics, governments and non-government organizations

Part V: Movies and literature