Empirical Essays in Artificial Intelligence Ethics

Martins Martinho Bessa, A.C.

DOI
10.4233/uuid:f686a46e-d470-4fb5-9a6d-997501188dfa

Publication date
2022

Document Version
Final published version

Citation (APA)
Martins Martinho Bessa, A. C. (2022). *Empirical Essays in Artificial Intelligence Ethics*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:f686a46e-d470-4fb5-9a6d-997501188dfa

EMPIRICAL ESSAYS IN ARTIFICIAL INTELLIGENCE ETHICS

EMPIRICAL ESSAYS IN ARTIFICIAL INTELLIGENCE ETHICS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus Prof. dr. ir. T.H.J.J. van der Hagen,
Chair of the Board for Doctorates
to be defended publicly on
Thursday 14 July 2022 at 12:30 o'clock

by

Andreia Cristina MARTINS MARTINHO

Master of Arts in Bioethics
New York University, United States of America
Born in Madeira Island, Portugal

This dissertation has been approved by the promotors:

promotor: Prof. dr. ir. C.G. Chorus
promotor: Dr. ir. M. Kroesen

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. C.G. Chorus	Delft University of Technology, promotor
Dr. ir. M. Kroesen	Delft University of Technology, promotor
Independent members:	
Prof. dr. N.J.A. van Exel	Erasmus University Rotterdam
Prof. dr. ir. G.J.P.M. Houben	Delft University of Technology
Prof. dr. M.J. van den Hoven	Delft University of Technology
Prof. dr. ir. I.R. van de Poel	Delft University of Technology

Keywords: Artificial Intelligence, Ethics, Morality, Empirical Research

Printed by: Studio Ridderprint

Front & Back: Sandra Tukker - Studio Ridderprint

ISBN 978-94-6384-354-6

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

Success is liking yourself, liking what you do, and liking how you do it.

Maya Angelou

SUMMARY

As Artificial Intelligence (AI) becomes increasingly important in modern society, there is a pressing need to address the ethical issues associated with these technologies. AI Ethics is a necessary endeavor to capitalize on the benefits of AI while minimizing its risks. However, it faces important challenges related to normative urgency, multi-purpose nature of AI, and multitude of stakeholders operating in the AI space. This doctoral dissertation builds on the premise that empirical information is valuable for AI Ethics to address these challenges and realize its normative mandate. The main ambition is to make an empirical contribution that facilitates the reflective development of AI, which assists the communities operating in the AI space to engage in a critical reflection on AI.

To meet this ambition AI is conceptualized through a morality lens. In this dissertation *Ethical AI* emerges as a central element featuring three different dimensions: (i) data analysis tool in morality; (ii) system in morally charged context; and (iii) artificial moral agent. Each dimension is operationalized through empirical methods from the social sciences toolkit.

The first dimension of *Ethical AI* features AI as a data analysis tool in morality. Learning algorithms are increasingly used in various scientific domains but remain largely unexplored in morality. In this dissertation, a unsupervised learning algorithm (latent class cluster analysis) is used to gain further insights into Moral Foundations Theory (MFT). The MFT is a theoretical model of intuitive ethics that builds on the assumption that a small number of moral foundations related to different adaptive challenges emerge as intuitive reactions to social stimuli. Recent MFT research has explored the relationship between moral foundations and human behavior. However, these studies rely on self-scoring and social experiments, which have well-documented limitations. The study presented in chapter 2 expands this line of research by investigating the relationship between moral foundations in morally charged legal narratives and corresponding legal outcomes. Three different clusters were identified in the data, thus unraveling latent associations between the moral foundations mapped in the legal narratives and corresponding legal outcomes. This study contributes to MFT research by further exploring the relationship between moral foundations and behavior while also illustrating the use of learning algorithms to gain insights into morality.

The second dimension of *Ethical AI* features AI as a system operating in a morally charged context. Two different studies are presented in this dimension. In the first study, the AI is an autonomous vehicle (AV) (chapter 3). Much attention has been paid to the ethics of autonomous driving, yet little information is available about the ethical issues in focus by the AV industry. A systematic review was conducted to compare and contrast the normative narratives in scientific and industry communities. The overall conclusion that can be drawn is that industry and academia look at the ethics of AV technology through rather different lenses. While the scientific literature has been primarily preoccupied with considerations of abstract moral dilemmas (trolley problems), industry reports adopt a much more pragmatic,

technology-infused, and perhaps overly optimistic narrative when discussing the potential of so-called edge cases where accidents cannot be avoided and loss of life and damage need to be minimized. More specifically, the findings in this research suggest that: (i) given the plethora of ethical issues addressed in the reports, autonomous driving companies seem to be aware of and engaged in the ethics of autonomous driving technology; (ii) scientific literature and industry reports prioritize safety and cybersecurity; (iii) scientific and industry communities agree that AVs will not eliminate the risk of accidents; (iv) scientific literature on AV technology ethics is dominated by discussions about the trolley problem; (v) moral dilemmas resembling trolley cases are not addressed in industry reports but there are nuanced allusions that unravel underlying concerns about these extreme traffic situations; (vi) autonomous driving companies have different approaches with respect to the authority of remote operators; and (vii) companies seem invested in a lowest liability risk design strategy relying on rules and regulations, expedite investigations, and crash/collision avoidance algorithms.

In the second study, the AI is a system operating in healthcare (chapter 4). Implementing these technologies in the health domain is promising, but it also raises important ethical issues. In this study, medical doctors were surveyed about the ethics of Health AI. Four main perspectives have emerged in the data representing different views about this topic. The first perspective (*AI is a helpful tool: Let physicians do what they were trained for*) highlights the efficiency associated with automation, which will allow doctors to have the time to focus on expanding their medical knowledge and skills. The second perspective (*Rules & Regulations are crucial: Private companies only think about money*) shows strong distrust in private tech companies and emphasizes the need for regulatory oversight. The third perspective (*Ethics is enough: Private companies can be trusted*) puts more trust in private tech companies and maintains that ethics is sufficient to ground these corporations. The fourth perspective (*Explainable AI tools: Learning is necessary and inevitable*) emphasizes the importance of explainability of AI tools in order to ensure that doctors are engaged in technological progress. Each perspective provides valuable and often contrasting insights about ethical issues that should be operationalized and accounted for in the design and development of Health AI.

The third dimension of *Ethical AI* features AI as an artificial moral agent (AMA). The development of AMAs is central to the Machine Ethics project, and it is quite controversial. Two different studies are presented in this dimension. In the first study AI Ethics scholars were surveyed on key issues of artificial morality (chapter 5). Five main perspectives about AMAs were empirically identified and discussed:(i) *Machine Ethics: The Way Forward*; (ii) *Ethical Verification: Safe & Sufficient*; (iii) *Morally Uncertain Machines: Human Values to Avoid Moral Dystopia*; (iv) *Human Exceptionalism: Machines Cannot Moralize*; (v) *Machine Objectivism: Machines as Superior Moral Agents*. The diverse perspectives about artificial morality identified in this study unravel the need for grounding and practicality in Machine Ethics. The second study builds on the notion that moral uncertainty is paramount to the artificial morality endeavor. A theoretical metanormative framework for equipping AI with moral uncertainty is operationalized through a latent class choice model. This study provides an empirical illustration of moral uncertainty thus opening avenues for further research on the meaning and practical implications of moral uncertainty in artificial decision-making (chapter 6).

This doctoral research illustrates the potential of using an empirical approach, along the

spectrum of empirical ethics and morally relevant social sciences, in morality without compromising the diversity of thought and normative edge. The studies featured in this dissertation provide important systematic and operational insights into the normative issues of AI. These insights should contribute to a critical development of AI. However, there is no conclusive evidence or measurement of such contribution. This is a common limitation in exploratory research, which is aggravated by the complex nature of this subject matter.

SAMENVATTING

Naarmate Kunstmatige Intelligentie (AI) steeds belangrijker wordt in de moderne samenleving, is er een dringende behoefte aan AI-ethiek om de ethische kwesties die verband houden met AI te adresseren. AI-ethiek is een noodzakelijk streven om de voordelen van AI te benutten en tegelijkertijd de risico's ervan te minimaliseren. Dit streven gaat gepaard met belangrijke uitdagingen die verband houden met de normatieve urgentie, de multifunctionele aard van AI, en de veelheid aan belanghebbenden die actief zijn in AI. Dit proefschrift gaat uit van de premisse dat empirische informatie waardevol is voor AI-Ethiek om deze uitdagingen aan te gaan en om AI's normatieve mandaat te realiseren. De belangrijkste ambitie van dit proefschrift is om een empirische bijdrage te leveren die de reflectieve ontwikkeling van AI vergemakkelijkt, en die de gemeenschappen die actief zijn in AI helpt om een kritische reflectie over AI aan te gaan.

Om aan deze ambitie te voldoen wordt AI geconceptualiseerd door een morele lens. In deze dissertatie komt *Ethische AI* naar voren als een centraal element met drie verschillende dimensies: (i) als data-analyse-instrument in moraliteit; (ii) als systeem in moreel geladen contexten; en (iii) als een kunstmatige morele 'agent'. Elke dimensie wordt geoperationaliseerd aan de hand van empirische methoden uit de sociale wetenschappen.

De eerste dimensie van *Ethische AI* kenmerkt AI als een data-analyse-instrument in moraliteit. Lerende algoritmen worden steeds meer gebruikt in verschillende wetenschappelijke domeinen, maar blijven grotendeels onontgonnen in morele vraagstukken. In dit proefschrift wordt een niet gesuperviseerd leeralgoritme (latent class cluster analysis) gebruikt om meer inzicht te krijgen in de Moral Foundations Theory (MFT). De MFT is een theoretisch model van intuïtieve ethiek dat aanneemt dat een klein aantal morele grondslagen in de loop van de menselijke evolutie (met alle adaptieve uitdagingen daarbinnen) zijn ontstaan als intuïtieve reacties op sociale stimuli. Recent MFT-onderzoek heeft de relatie tussen deze veronderstelde overal geldende en tijdloze morele grondslagen en het menselijk gedrag onderzocht. Deze studies zijn echter gebaseerd op zelf-scoring en sociale experimenten, die beperkingen hebben die reeds vaak zijn gedocumenteerd. De studie in hoofdstuk 2 breidt deze onderzoekslijn uit door de relatie te onderzoeken tussen morele grondslagen in moreel geladen juridische verhalen en overeenkomstige juridische uitkomsten met een niet gesuperviseerd leeralgoritme. Drie verschillende clusters werden geïdentificeerd in de data, waarmee latente associaties tussen de morele fundamenten in de juridische verhalen en corresponderende juridische uitkomsten werden ontrafeld. Deze studie draagt bij aan het MFT onderzoek door de relatie tussen morele grondslagen en gedrag verder te onderzoeken en illustreert tevens het gebruik van lerende algoritmen om inzicht te krijgen in moraliteit.

De tweede dimensie van *Ethische AI* kenmerkt AI als een systeem dat opereert in een moreel geladen context. In deze dimensie worden twee verschillende studies gepresenteerd. In de eerste studie is de AI een autonoom voertuig (AV) (hoofdstuk 3). Er is veel aandacht besteed aan de ethiek van autonoom rijden, maar er is weinig informatie beschikbaar over

de ethische kwesties waar de AV-industrie zich op richt. Een systematische review werd uitgevoerd om de normatieve verhalen in wetenschappelijke en industriële gemeenschappen te vergelijken en te contrasteren. De algemene conclusie die hieruit kan worden getrokken is dat de industrie en de academische wereld de ethiek van de AV-technologie door nogal verschillende lenzen bekijken. Terwijl de wetenschappelijke literatuur zich in de eerste plaats bezighoudt met beschouwingen over abstracte morele dilemma's (trolleyproblemen), hanteren de rapporten van de industrie een veel pragmatischer, technologie-gedreven, en misschien overdreven optimistisch verhaallijn wanneer ze het potentieel bespreken van zogenaamde randgevallen waar ongevallen niet kunnen worden vermeden en verlies van levens en schade tot een minimum moeten worden beperkt. Meer specifiek suggereren de bevindingen in dit onderzoek dat: (i) de AV-industrie lijkt zich bewust van en betrokken bij de ethiek van autonoom rijdende technologie gezien de overvloed aan ethische kwesties die in de rapporten aan bod komen; (ii) zowel wetenschappelijke literatuur als rapporten uit de industrie geven prioriteit aan veiligheid en cyberveiligheid; (iii) wetenschappelijke en industriële gemeenschappen zijn het erover eens dat AV's het risico op ongevallen niet zullen elimineren; (iv) wetenschappelijke literatuur over ethiek van AV-technologie wordt gedomineerd door discussies over het trolleyprobleem; (v) morele dilemma's die lijken op trolleyproblemen worden niet behandeld in rapporten van de industrie, maar er zijn genuanceerde toespelingen die onderliggende bezorgdheid over deze extreme verkeerssituaties laten zien; (vi) de industrie heeft verschillende visies op de autoriteit van degene(n) die van afstand op AV's kunnen ingrijpen; en (vii) bedrijven lijken te investeren in een ontwerpstrategie met een zo laag mogelijk aansprakelijkheidsrisico en vertrouwen op regels en voorschriften, snelle onderzoeken en algoritmen om botsingen en ongevallen te vermijden.

In de tweede studie is de AI een systeem dat actief is in de gezondheidszorg (hoofdstuk 4). De toepassing van deze technologieën in de gezondheidszorg is veelbelovend, maar werpt ook belangrijke ethische vragen op. In deze studie zijn medici ondervraagd over de ethiek van Health AI. Vier belangrijke perspectieven zijn naar voren gekomen in de gegevens die verschillende opvattingen over dit onderwerp vertegenwoordigen. Het eerste perspectief (*AI is een nuttig instrument: laat artsen doen waarvoor ze zijn opgeleid*) benadrukt de efficiëntie die gepaard gaat met automatisering waardoor artsen de tijd zullen hebben om zich te concentreren op het uitbreiden van hun medische kennis en vaardigheden. Het tweede perspectief (*Regels & voorschriften zijn cruciaal: privébedrijven denken alleen aan geld*) toont een sterk wantrouwen in technologiebedrijven en benadrukt de behoefte aan regulerend toezicht. Het derde perspectief (*Ethiek is genoeg: private bedrijven zijn te vertrouwen*) geeft meer vertrouwen aan technologiebedrijven en stelt dat de ethiek binnen die bedrijven voldoende is. Het vierde perspectief (*Uitlegbare AI-tools: leren is noodzakelijk en onvermijdelijk*) benadrukt het belang van uitlegbaarheid van AI-tools om ervoor te zorgen dat artsen betrokken zijn bij technologische vooruitgang. Elk perspectief biedt waardevolle en vaak contrasterende inzichten over ethische kwesties die kunnen worden geoperationaliseerd en verantwoord in het ontwerp en de ontwikkeling van Health AI.

De derde dimensie van Ethische AI kenmerkt AI als een kunstmatige morele agent ('artificial moral agent' - AMA). De ontwikkeling van AMA's staat centraal in het Machine Ethics project, en is behoorlijk controversieel. In deze dimensie zijn twee verschillende studies uitgevoerd. In de eerste studie werden 'AI Ethics'- wetenschappers ondervraagd

over de belangrijkste kwesties van kunstmatige moraliteit (hoofdstuk 5). Vijf belangrijke perspectieven over AMA's werden empirisch geïdentificeerd en besproken: (i) *De Ethiek van Machines: De Weg Voorwaarts*; (ii) *Ethische Verificatie: Veilig & Voldoende*; (iii) *Moreel Onzekere Machines: Menselijke Waarden om Morele Dystopie te vermijden*; (iv) *Menselijk Exceptionalisme: Machines kunnen niet moraliseren*; (v) *Objectieve Machines: Machines als Superieure Morele Agenten.* De verschillende perspectieven over kunstmatige moraliteit die in deze studie zijn geïdentificeerd, laten de behoefte aan onderbouwing in en bruikbaarheid van Machine Ethiek zien. De tweede studie bouwt voort op het idee dat morele onzekerheid van groot belang is om te komen tot kunstmatige moraliteit. In deze studie wordt -door middel van een latente klasse keuzemodel- een theoretisch metanormatief raamwerk geoperationaliseerd om een AI uit te rusten met morele onzekerheid. Deze studie biedt een empirische illustratie van morele onzekerheid en opent zo wegen voor verder onderzoek naar de betekenis en praktische implicaties van morele onzekerheid in kunstmatige besluitvorming (hoofdstuk 6).

Dit onderzoek illustreert het potentieel van het gebruik van een empirische benadering in het brede spectrum van empirische ethiek en moreel relevante sociale wetenschappen zonder afbreuk te willen doen aan de diversiteit in het denken over ethiek. De studies in dit proefschrift bieden belangrijke systematische en operationele inzichten in de normatieve kwesties van AI. Deze inzichten kunnen bijdragen aan een kritische ontwikkeling van AI. Er is echter momenteel geen sluitend bewijs voor of concreet inzicht in een dergelijke bijdrage. Dit is een gebruikelijke beperking van verkennend onderzoek, die nog wordt verergerd door de complexe aard van dit onderwerp.

ACKNOWLEDGEMENTS

I have always been interested in observing patterns. Still to this day, my mother is mortified when recalling three-year-old me walking through the church only to inquire a woman about why she would wear the same dress every Sunday. As the woman made some excuse for her *faux pas*, my mother realized she was in trouble as my observation could not be dismissed as gibberish.

To a great extent, I am still clinging to that interest in behavioral patterns, as I use sophisticated statistical tools to identify clusters in the convoluted space of Artificial Intelligence Ethics. This interest was important to pursue my doctoral research, but much more is needed to complete a Ph.D. Every doctoral candidate needs an opportunity and support.

I thank my supervisor Caspar Chorus for the opportunity to pursue a Ph.D. in his research group. For several years I had the ambition to do a Ph.D. and I thank you for believing in me. I am truly passionate about research and I hope I can honor this opportunity by fostering a scientific career.

Throughout my Ph.D. I received enormous support from my family, supervisors, and friends. I thank my parents, Manuel Martinho and Maria José Martinho, for the freedom to dream beyond the sharp mountains of Madeira Island. As a child, I may have caused you some socially embarrassing situations, but you always chose to see my observations as a sign of intelligence rather than inconvenience. I often think how wonderful it would be to somehow go back in time and give you the same opportunities that I had. I thank my little sister Joana Martinho for being the ray of light in my life since the day she was born. I thank my brilliant husband Miguel Bessa - you inspire me every day.

I also want to express gratitude and appreciation to my supervisors. Caspar and Maarten Kroesen patiently and diligently helped me become a researcher. There are a lot of things to write about Caspar. I am amazed by his organization, writing, and conceptualization skills. But perhaps what I most admire in him is that he leads by example - he sets the bar high but somehow his own bar seems to be higher. I am also very grateful to Maarten. He came into the project later and was a suave element between two big personalities. I thank him for being friendly, kind, and always willing to help whenever I got stuck in the statistical modeling.

Finally, I acknowledge the support from my friends. I was fortunate to share my Ph.D. experience with a diverse group of researchers (Teodóra Szép, Tanzhe Tang, Tom van den Berg, Nicolas Cointe) who fostered a positive and supportive research environment. I also thank my friends from Madeira island and my friends from the University of Coimbra - no matter how far I am, you always make me feel close. In particular, I thank my friend Cristina Pereira, who was a true mentor during my Ph.D.

CONTENTS

1

INTRODUCTION

The renewed interest in Artificial Intelligence (AI) is a moment of truth for AI Ethics. No longer confined to a small space of moral debate, it is now on the agenda of governments, private companies, and non-governmental organizations. In this larger space, there are novel challenges for AI Ethics to realize its normative mandate. The fast development of AI demands urgent guidance, contrasting with the traditionally slow normative process. Since AI is a multi-purpose technology currently being developed in different societal fields, normative work needs to be contextualized. The multitude of AI stakeholders requires clear communication between the normative and technology communities operating in the AI space. This doctoral dissertation builds on the premise that empirical work is valuable for AI Ethics to address these challenges, thus facilitating its normative endeavor. The foundations of the dissertation are presented in this introductory chapter featuring a research background (section 1.1), research aim and questions (section 1.2), research methodology (section 1.3), and structure of the dissertation (section 1.4).

1.1. RESEARCH BACKGROUND

1.1.1. OVERVIEW

The starting point of this doctoral dissertation is a systematic research background. The current data-driven paradigm in AI is briefly characterized. It is emphasized that AI has the potential to achieve social good but also harmful, unfair, or obscure outcomes (sub-section 1.1.2). By contrasting these potential outcomes, the stage is set for AI Ethics as a necessary endeavor to capitalize on the benefits of AI while minimizing its risks. A historical contextualization of the normative narratives about AI is provided in order to frame the ongoing debates (sub-section 1.1.3) and challenges faced by AI Ethics (sub-section 1.1.4). It is premised that empirical work is valuable for AI Ethics to address those challenges. However, it is also acknowledged that the use of empirical research in the normative domain is controversial. Different combinations of empirical and normative research are outlined only to frame this work along the spectrum of empirical ethics and morally relevant social sciences (sub-section 1.1.5). The empirical research featured in the dissertation is further contextualized in AI Ethics (sub-section 1.1.6) with respect to the different dimensions of

Ethical AI (sub-section 1.1.7).

1.1.2. THE DATA-DRIVEN PARADIGM IN AI

AI is everywhere. Unprecedented availability of data resulting from the widespread use of wireless network technologies [1, 2], along with algorithmic innovation [3, 4], and robust computing power [5, 6], has prompted the emergence of a data-driven paradigm in AI. The modern AI feeds on data to learn solutions for problems of Science [7–13], Technology [14–21], Arts [22–25], War [26–28], and even Love [29]. The predominance of AI in the scientific arena is well illustrated by the impressive number of AI publications across different fields (Figure 1.1) [1] as well as the number of AI journals launched recently by the largest commercial publishers (Figure 1.2) [2].

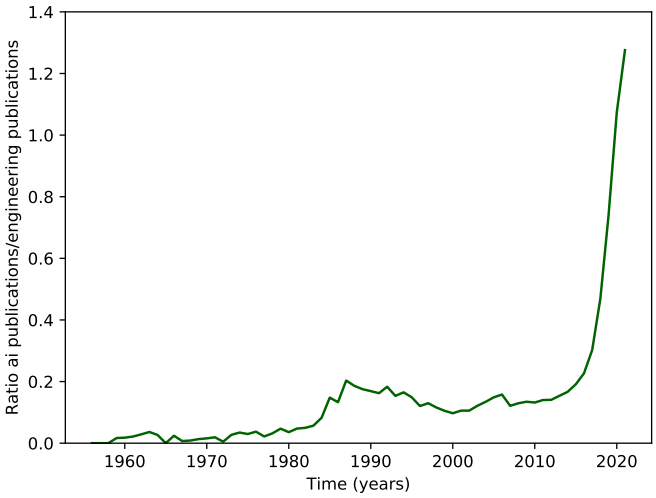
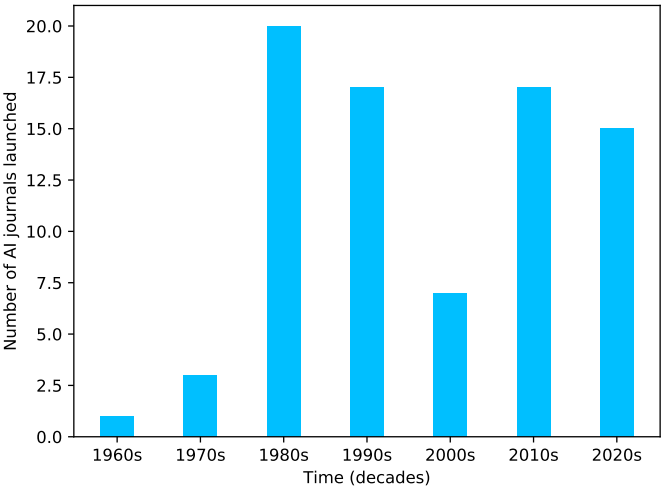


Figure 1.1: AI research throughout the years (1956-2021).

[1] The AI research is computed as a ratio of AI publications over Engineering publications. These publications were identified in Web of Science through the use of keywords "Artificial Intelligence", "Machine Learning", and "Deep Learning" (AI publications), and "Engineering" (Engineering publications), within the timeline 1956-2021. The ratio of AI publications over Engineering publications was computed for each year. The plotted data in Figure 1.1 indicates that AI publications have now surpassed Engineering publications, which is speculated to be a result of an increased interest of AI not only in Engineering but also in other scientific domains.

[2] The trends related to AI journals were identified through the inspection of the databases of the largest publishers (Elsevier, Springer, Wiley, Taylor & Francis, and Sage [30]). The selection of journals resulted from the use of keywords "Artificial Intelligence", "Intelligence", "AI", "Intelligent", "Expert Systems", "Knowledge based", 'Knowledge', "Machine Learning", "Deep Learning", "Neural Networks", "Pattern Recognition", and "Autonomous" in the databases of publishers. Subsequently, the launch date of each AI journal was registered and the journals were clustered into different decades (1960s - 2020s). It is noted that the databases only feature existing journals, thus leaving out journals that may have been launched in the past but are no longer operating. The plotted data indicates an impressive number of journals launched in the first two years of the 2020s.



H

Figure 1.2: AI journals launched throughout the decades (1960s-2020s) by publishers Elsevier, Springer, Wiley, Taylor & Francis, and Sage (2020s decade features only years 2020 and 2021).

Within this paradigm, the approach to intelligence relies heavily on sub-symbolic algorithms that learn and improve through experience without the need for explicit programming [31]. Symbolic approaches to AI, which attempt to represent cognitive functions such as thinking, learning, and problem-solving through the explicit programming of symbolic reasoning and logic [32–34], remain relevant in many research and industry communities [35–41]. Yet, the staple of the current data-driven AI paradigm are sub-symbolic learning algorithms that thrive on data and are particularly well suited to solve domain-specific problems [33].

Formally, Machine Learning (ML) algorithms are defined as being capable of learning from experience E with respect to some class of tasks T, as long as its performance at tasks in T improves with experience E (as measured by a performance metric P) [42]. In practice, the class of tasks to be learned, abundance and quality of data available for training, as well as the choice of ML algorithm, can have dramatic influence on the performance. Some applications are effectively addressed by simpler algorithms such as Support-Vector Machines [43], Naive Bayes [44], and conventional Artificial Neural Networks [45], while others require deep architectures with many hidden layers, involving building blocks such as convolutional filters [46] and recurrent units [45, 47].

In recent years learning algorithms have become a core part of AI Systems that have permeated into the daily life of developed societies. For now, these systems operate mainly as supportive technologies. AI-powered mobile and web-based applications assist users in a wide range of tasks such as planning (e.g. weather and traffic forecast) [48], monitoring health [49–52], and navigating legal systems [53–55]. Chatbots use AI to engage in natural language conversations and are increasingly used across many industries and sectors to

facilitate the interaction between users and organizations [56–65]. Autonomous robots are machines that rely on AI to perform varied tasks such as household cleaning [66, 67], providing care for the elderly [68, 69], and driving [15, 70, 71].

Autonomous driving technology provides a compelling case study of the societal impact of AI. Research in this field is motivated by the large number of traffic accidents caused by human errors [72]. Building on the data-driven AI breakthrough, significant advances were made in the automation of the driving task in the last decade [73–75], with the goal of reducing the number of traffic accidents. Today many of the vehicles available in the market feature low levels of automation that support the driving task, but AI opened the door for higher automation. These AI-powered autonomous driving systems operate as replacement technologies that take over the driving task. Ultimately, if full autonomy is achieved and deployed in the market without operational design domain restrictions, these technologies may become truly disruptive [76].

There is an enormous potential for AI technologies to enhance human well-being by enabling key sustainable development goals [77, 78]. In transportation and sustainable cities, electric autonomous vehicles may play an important role in enabling sustainable low-carbon cities amid the current climate emergency and low-carbon transition [76, 77]. In the healthcare space, AI may be used in low and middle-income countries to mitigate the shortage of healthcare resources, namely by improving the reliability of parasite detection tests for the diagnosis of communicable diseases, thus promoting health and well-being [77, 79–82]. On peace and justice, AI has the potential to promote the rule of law and combat organized crime through predictive policing, and automatic fraud detection algorithms [77, 83–85]. AI will not solve complex social phenomena, but it has the potential to promote good social outcomes within those spaces.

Despite the potential for AI to achieve social good, there are also serious concerns that these technologies may cause harmful, unfair, or obscure outcomes [84, 86–93]. Provided that AI technologies are still very much in development, there are significant safety concerns. In the past years several incidents and accidents involving AVs were reported [94–96]. Most of the incidents were minor rear-ended crashes [95, 97] but there are also reports of tragic fatalities. In 2018 a man lost his life while riding an AV in California that crashed into a white truck not detected by the sensors, and, in the same year, a pedestrian woman also lost her life as she was tragically hit by an AV in Arizona [97–99].

There are also concerns about AI Systems producing biased outputs [93]. The datasets that sustain AI technologies often feature errors, ill judgments, and prejudices which can enter into the innovation lifecycle and create bias [87]. A few illustrative cases are so often used in the literature that have now become textbook examples of AI bias: the Amazon ML recruitment algorithm that penalized female applicants [100, 101]; the dermatologic AI applications that are not as effective in populations poorly represented in the medical datasets [102]; and the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) ML algorithm that was racially biased when calculating recidivism risk for parole and bail in the U.S justice system [86, 93, 103]. There are also concerns about AI Systems making obscure decisions. In the credit scoring literature, it is reported that ML models have the potential to replace traditional regression models due to their higher predictive power. However, such models are generally unable to provide explanations about their predictions [104, 105]. Given that credit scoring models used by financial institutions

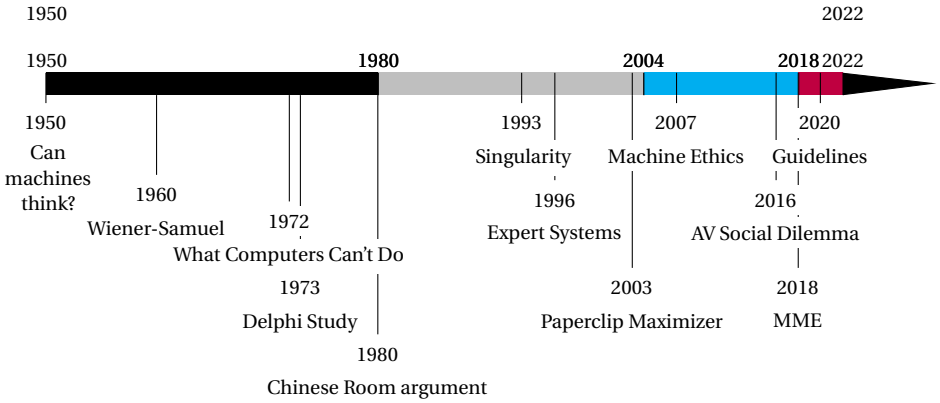
are required to be transparent, the obscurity of advanced ML models represents a lost opportunity in the financial industry.

As the presence of AI in modern society becomes increasingly important, and its capacity to do good and harm becomes clear, there is a pressing need to address the ethical issues associated with these technologies. AI Ethics is a necessary endeavor to capitalize on the benefits of AI while minimizing its risks. The scientific community has responded to this call, and, in recent years, an extraordinary amount of work has been developed within the realm of AI Ethics.

1.1.3. MAPPING AI ETHICS THROUGHOUT THE YEARS

Normative discussions about AI have run parallel to its development throughout the years (Figure 1.3). These discussions reflect the achievements, failures, and societal anxieties associated with these technologies. Some of the issues explored in the past, related to explainability, accountability, or privacy, remain relevant today.

Figure 1.3: Milestones: AI Ethics debate throughout the years.



AI ETHICS: THE EARLY DAYS

In the early days of AI, the focus was on getting computers to do things that would be regarded as intelligent if done by humans [106]. The philosophical discussions that took place around this time were not so much about morality but rather about epistemological questions concerning minds and machines: *Can machines think?*, *Can machines be intelligent?*, *Can an artificial brain potentially outperform a human brain?* [106, 107]. These foundational questions would continue to be explored throughout the following decades, but the philosophical debate became more diverse as AI evolved.

Some moral considerations surfaced in the literature of the 1960s [108]. At this time, reckless claims were made regarding the potential of AI to copy and perhaps even recreate the entire workings of the human brain within a short timeline [106]. Those claims prompted concerns about AI technology. The Wiener-Samuel debate illustrates early discussions about the potential for machines to threaten humankind. Norbert Wiener famously wrote

that *machines can and do transcend some of the limitations of their designers, and in doing so they may be both effective and dangerous* [109]. Arthur Samuel disagreed on the grounds that machines could not do anything unless instructed by humans [110]. He conceded, however, that *projected machines of the so-called "neural net" type* could be an exception to the deterministic nature of machines. Since their internal connections would be unknown, the precise behavior of the nets would be unpredictable and, therefore, potentially dangerous [110]. Ahead of his time, Samuel was alluding to what is known today as the problem of *explainability* associated with sub-symbolic AI algorithms.

The ethical and societal implications of AI were initially explored in the 1970s. This was a period of great disenchantment with AI, as the technology had failed to meet expectations and deliver the promised results [106]. In addition to the technical frustrations and lack of funding, important philosophical work challenging the claims of AI was published around this time. Hubert Dreyfus published *What Computers Can't do*, in which he argued that machines could not display higher mental functions through the use of symbolic representations [111, 112]. John Searle came up with his Chinese room argument to show that a computer could not be said to understand the symbols with which it communicates [113]. Eventually, the claims about AI were toned down. Many researchers followed the lead of John McCarthy to develop AI programs limited to a particular domain of knowledge with practical applications in industry [106, 114].

Around this period, a Delphi study was conducted among experts in the AI field with the purpose of understanding both the capabilities and limitations of machine intelligence and its potential impact on society [115]. This study accurately predicted the use of AI in mundane activities such as domestic chores, entertainment, or weather forecast and addressed ethical issues related to safety, privacy, and jobs displacement. It was also cautioned that there was a need to formalize algorithmically some of the ethical and empirical rules and trade-offs that society observes implicitly [115]: *It may be necessary for a robot to have available rules which dictate how to trade-off life for property, e.g., when is it allowable to wreck an automobile to avoid killing an animal?* [115]. This sort of moral dilemma would eventually become the focal point of AI Ethics decades later.

AI ETHICS: ACCOUNTABILITY OF EXPERT SYSTEMS, SINGULARITY, & ASIMOV

Following the lead of McCarthy, throughout the 1980s and 1990s, the AI community focused on the development of expert systems. These computer systems were designed to solve complex reasoning problems in particular domains at the level of performance of human experts [116]. The normative discussions during the heyday of expert systems were residual, although some scholars explored issues of moral responsibility and accountability associated with these systems [117–119]. Expert systems were developed to operate in different domains, such as Medicine and Law [120]. However, they faced many challenges, mainly because the medical or legal daily practice can not be reduced to a closed set of rules [121, 122].

Despite only moderate success of Expert Systems, there was an ongoing conversation about technological singularity, i.e., a hypothetical context where super-intelligent machines design and produce even more super-intelligent machines [106, 123–125]. The *Paperclip Maximizer* introduced by Nick Bostrom, which describes the existential risk that superintelligence may pose to human beings when programmed to pursue seemingly harmless goals [125], is yet another illustration of the concerns over the potential of machines to threaten

humankind. Contrasting with those early discussions in the 1960s, this time, the scientific community was interested in finding solutions for the problem of complex systems lacking morality [126]. In the absence of a moral theory, the three laws of robotics, introduced by science fiction writer Isaac Asimov, became a fixture in scientific publications about the ethics of machines and AI [106].

AI ETHICS: MACHINE ETHICS, TROLLEY PROBLEMS, & EMPIRICAL ETHICS

Concerns about the moral behavior of increasingly intelligent and autonomous machines led to the emergence of Machine Ethics. This research field aims at equipping machines with ethical reasoning to ensure that their behavior towards humans and other machines is ethically acceptable [127–131]. The core idea is that sensitivity to ethics should be integral to the software of machines in order to facilitate their ethical use [129].

Unlike traditional philosophy of technology, which was primarily reactive, and modern philosophy of technology, which is proactive in raising awareness of the values designers bring to the technology design process, Machine Ethics went one step further, seeking to build ethical decision-making capacities directly into the machines [129, 132].

The need for artificial morality was reinforced with the advent of autonomous driving. The novel AMA was an AV in an extreme traffic situation. Upon a string of high profile publications [133, 134], the AV trolley problem, which had been timidly addressed in the 1970s Delphi study, was heavily featured in the scientific and popular literature [72, 98, 135–139]. The discussions revolved around the practical relevance of the AV moral dilemma [140–142], the merits of using different ethical frameworks as control algorithms for AVs [143–148], and the moral preferences and societal expectations about the ethics to be encoded in AVs [133, 134, 149, 150]. It is not unlikely that one day the scientific community will disapprove of the amount of attention paid to the AV trolley. However, it is indisputable that such attention played a crucial role in raising awareness about AI Ethics.

AI ETHICS: GUIDELINES & PRINCIPLES

Increased awareness about AI Ethics led to a proliferation of soft governance mechanisms for the ethical development and deployment of AI [151, 152]. Many organizations have launched a wide range of initiatives, such as codes, guidelines, frameworks, and policy strategies, to establish ethical principles for the adoption of socially beneficial AI [153]. These initiatives brought the focus of the AI Ethics debates to a common set of issues and principles. Although varying in terminology, it is reported that the different guidelines broadly converge around five principles: (i) beneficence, (ii) non-maleficence, (iii) autonomy, (iv) justice, and (v) explicability [151, 153, 154]. Ethically-aligned AI should therefore be beneficial to people and the environment (beneficence); robust and secure (non-maleficence); respectful of human values (autonomy); fair (justice); and explainable and accountable (explicability). These principles closely resemble the four classic principles in medical ethics [155]. The association is convenient since medical ethics is historically the most prominent and well-studied approach to applied ethics, however it may not be warranted due to important differences between the two fields [156] [3].

The development of these largely principle-based documents has been an important and necessary phase in the evolution of AI governance [157], as there is now an outline of a

[3] A proper account of the differences between the Medical and AI Ethics fields, which may hinder the impact of the Medical Ethics principles in the domain of AI, is provided by [156].

shared normative foundation about AI. Unlike classical principles in medical ethics, which provide a common ground that professionals can refer to in case of conflict, the same principles construed within the AI Ethics domain seem more intended to embed normative considerations in the design and governance of technology [156]. However, **because these principles remain abstract and are not translated to practice, the AI Ethics endeavor may be falling short of meeting its goals of providing normative guidance for the design and deployment of algorithms** [152, 158].

A poor translation of these principles and guidelines into practices leaves room for unethical behaviors such as *ethics shopping*, i.e., mixing and matching ethical principles from different sources to justify some pre-existing behavior [159]; *ethics bluewashing*, i.e., making unsubstantiated or misleading claims about the efforts and resources allocated to address ethical problems associated with AI; *ethics dumping*, i.e., exporting research activities about digital processes, products, services, or other solutions, in other contexts or places in ways that would be ethically unacceptable in the context or place of origin and importing the outcomes of such unethical research activities[159]; *ethics shirking*, i.e., doing increasingly less ethical work in a given context, thus lowering the overall level of ethics engagement [159]; and *ethics lobbying*, i.e., exploiting ethics to delay or avoid good and necessary legislation [159]. Concerns about these unethical behaviors related to AI are markedly present in the recent and more practical work developed in AI Ethics.

AI ETHICS TODAY

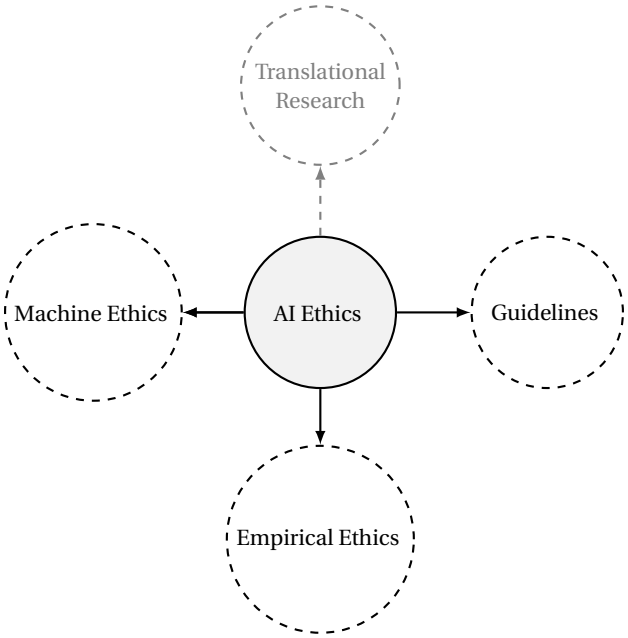
To a great extent, the normative work developed in previous decades has shaped the field of AI Ethics. Research carried out in AI Ethics ranges from (i) reflections and practical work on how ethical principles can be implemented in decision routines of autonomous machines (Machine Ethics) [146, 160–162]; (ii) empirical analysis on how moral dilemmas are solved (Empirical AI Ethics) [133, 134]; and (iii) comprehensive AI principles and guidelines (Principles & Guidelines)[151].

Despite the richness of the AI Ethics work, a growing body of the literature has declared that it is failing to realize its normative endeavors. There is little evidence on the plausibility of the Machine Ethics project, as AMAs remain, for now, proofs of concept and lab prototypes [162]. The empirical AI Ethics studies have provided rich information about the societal expectations and preferences in moral situations involving AI. However, some of these studies, such as the Moral Machine Experiment (MME), have been criticized for their stylized and unrealistic premises (moral dilemmas) [134, 163]. The impact of the AI Ethics guidelines developed to promote ethical practices in AI is also disputed. There is a mistrust that organizations, particularly those involved in the development of AI technologies, will implement ethical practices voluntarily [164]. When Ethics is integrated into organizations, there are concerns that it is used merely as a marketing strategy with little impact when it comes to decisions made in the AI domain [152, 165, 166]. A study that surveyed AI practitioners about their perceived impact of AI Ethics guidelines reported that the effectiveness of such guidelines or ethical codes is almost zero and that they do not change the behavior of professionals from the tech community [165].

The translational work aiming at producing tools or methods for implementing ethics into practice has also been scrutinized. There is little evidence of the impact of existing translational tools on the governance of AI [91]. Morley et al. identified the tools and methods already available to guide AI practitioners on core issues of AI Ethics and plotted

these methods and tools in a typology, matching them to ethical principles (beneficence, non-maleficence, autonomy, justice, and explicability) to stages in the algorithm development pipeline. They reported that numerous tools and methodologies exist to assist practitioners in realizing Ethical AI, but the vast majority are severely limited in terms of usability [91].

Figure 1.4: Graphical representation of AI Ethics.



Poor guidance on how to produce Ethical AI may result in significant societal opportunity costs [91]. AI is a multi-purpose technology with the potential to improve human well-being. However, consumers may be discouraged from adopting these technologies if the costs of ethical mistakes outweigh the benefits of ethical success. Weak public acceptance means slimmer chances for AI to meet its potential. Resorting once again to the archetypal AV case, it is clear that the ethical trade-offs in traffic situations that entail distribution of risk will need to be resolved for this technology to be accepted by the public, thus meeting its potential to save lives [133, 163, 167].

The so-called failure of AI Ethics to realize its normative mandate raises unsettling concerns about the feasibility of this endeavor. Some of those concerns may be unwarranted as stemming from a misconception about Ethics. It is not expected that AI Ethics provides policies or regulations but rather a operational normative framework. Past experience with other applied fields of Ethics (e.g., medical ethics) shows that it is possible to operationalize Ethics successfully, or at least abstract ethical principles, thus lending credence to the AI Ethics effort [158]. It is hypothesized that the current struggles of AI Ethics may be rooted in the challenges that have emerged in the current data-driven paradigm in AI and that empirical work may be valuable in overcoming those challenges.

1.1.4. NOVEL CHALLENGES FOR AI ETHICS: PRACTICAL NORMATIVE GUIDANCE

AI Ethics is no longer confined to a small space of moral debate. In the current data-driven paradigm, an extraordinary amount of work has been developed by the scientific community, governments, private companies, and non-governmental organizations (e.g., guidelines, codes of conduct, principles). Operating in a larger space, AI Ethics faces novel challenges related to the urgency of normative guidance, multi-purpose nature of AI, and multitude of stakeholders.

Ethics explores matters of right and wrong that reflect the moral spectrum about a particular topic. The normative debates welcome disagreement, speculation, and abstraction as a reflection of diversity of thought and serious ethical consideration [156]. The purpose is to slowly build a robust, rich, and diverse theoretical normative foundation that will eventually lead to governance and regulatory mechanisms.

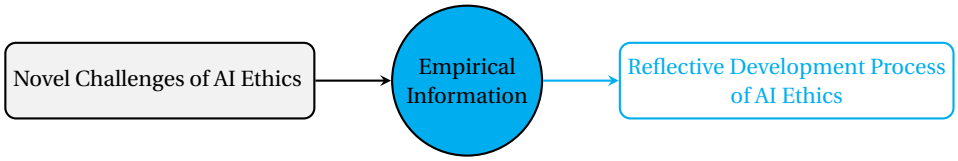
The AI Ethics work developed in recent years meets the traditional expectations of the field in terms of richness, controversy, and diversity of thought. Nevertheless, given the current state of affairs in AI, characterized by the fast development and deployment of these technologies, there is an urgent need for practical and operational normative guidance.

Another challenge for AI Ethics is related to the fact that modern AI is a multi-purpose technology. Contrasting with other disruptive and ethically problematic technologies, such as nuclear transfer, induced pluripotent stem cells (iPS cells), or clustered regularly interspaced short palindromic repeats (CRISPR), that are somewhat self-contained in the life sciences domain, AI is currently applied to several societal domains, such as Transportation, Healthcare, or Justice. AI Ethics needs to consider how risks, conflicting rights and interests, and social preferences vary in different contexts [158].

Given the multi-purpose nature of AI, there are many stakeholders involved in the development of these technologies. These stakeholders, namely organizations such as Google, Amazon, Facebook, or DeepMind, have shown a strong interest in AI Ethics and have developed several guidelines [152]. Traditionally, ethicists have been hesitant about industry-led initiatives regarding Ethics and often dismiss these initiatives as attempts to shape the normative conversations to their own interests [91, 151]. However, it can also be hypothesized that these organizations produce guidelines because they have poor support from the Ethics communities. The Ethics work is rarely "ready to use" as it is often abstract, speculative, and intricate. A challenge for ethicists is to operationalize their work and communicate clearly and effectively with the other communities in the AI space.

This doctoral dissertation builds on the premise that there is value in using empirical information to address these challenges. **The main ambition is to make an empirical contribution that facilitates the AI Ethics reflective development process, which assists the communities operating in the AI space to engage in a critical reflection about AI** [158, 168, 169]. However, it is noted that the use of empirical work in the domain of Ethics is quite controversial. Different possible combinations of empirical and normative research are outlined below to frame the empirical work presented in this dissertation properly.

Figure 1.5: Graphical representation of the relation between the challenges of AI Ethics, empirical information, and reflective development.



1.1.5. ETHICS & EMPIRICAL RESEARCH

The relation between normative and empirical fields is not optimal. Pragmatic reasons related to the different nature and aims of these disciplines are part of the problem. Ethics is a prescriptive discipline focused on moral evaluation, conceptual clarification, and normative justification, whereas empirical sciences is a descriptive discipline focused on description, reconstruction, and analysis [170]. Historically, Ethics is modeled after the field of Philosophy to produce a rational and de-contextualized discourse. Accordingly, it tends to reject the empirical sciences amid the fear that sociological contextualization may lead to ethical relativism [170]. Important meta-ethical fallacies (*is-ought problem*; *naturalistic fallacy*; and *fact-value distinction*) [4] have provided theoretical ground for such rejection and are often used to dismiss the empirical work in the normative domain. [172].

Despite the strained relation between normative and empirical sciences, it is not uncommon for ethics to undergo an empirical turn in order to improve its realism and context specificity [170, 173, 174]. Such a turn was evident in the field of medical ethics. As a response to the critiques of foundationalism and as part of the attempt to operationalize its key principles, the focus of medical ethics shifted from purely normative to an approach that is more grounded in practical reality [170].

Broad conceptions of empirical ethics include empirical and philosophy-driven research. In *Empirically driven empirical ethics*, empirical claims describe or explain the world either to identify normative issues that arise in a particular domain of practice or to make sense of the relevant experiences, understandings, judgments, or intuitions of individuals concerning those issues. In *Philosophically driven empirical ethics*, normative principles and guidelines are operationalized with the aim of practical implementation. Both conceptions require a prescriptive element where some course of action is recommended. Still, this element is more relevant in the *empirically driven empirical ethics*, as it is distinctive of the social scientific endeavor of describing normative dimensions of practice [175].

Other more nuanced combinations of empirical and normative research have also been described in the literature [172, 175–177]. These combinations tend to feature empirical and normative research as *parallel*, *symbiotic*, or *integrated*. A *parallel* combination means that the disciplines have each their own specific task which cannot be reduced to the other

[4] The classical formulation of the *is-ought problem* is found in a passage of *A Treatise of Human Nature* by David Hume and relates to the fallacy of drawing an ought-conclusion from a set of is-premises [171]. The *naturalistic fallacy* by George Edward Moore (*Principia Ethica*) relates to the fallacy of identifying the predicate good with any natural or metaphysical predicate [171]. The *fact-value distinction* refers to different meta-ethical views about facts and values: (i) no statement or concept is irreducibly both evaluative or factual; (ii) evaluative discourse fails to have certain characteristics essential to paradigmatic realistic discourses like the scientific discourse; and (iii) scientific facts do not presuppose values (science is value free) [171].

[176, 177]. A *symbiotic* combination entails that both disciplines influence each other in various ways. Empirical research may be used solely to ensure the application of the prescriptive function of moral theories (moral theory → empirical data) or to improve moral theories (moral theory → empirical data → moral theory), or it can also be part of a more dynamic process in which the critical function of Ethics is applied to empirical data and simultaneously the moral theory is criticized by empirical data (moral theory ⇄ empirical data) [172]. A *integrated* combination may be said to be a new discipline with its premises, theories, topics, and methods to integrate empirical and normative research (moral theory ≈ empirical data) [176, 177] in which ethicists and descriptive scientists try to integrate moral theory and empirical data in order to reach a normative conclusion concerning a particular social practice [172].

The research featured in this dissertation shares the empirical ethics overarching aim of improving context specificity of normative work [173]. Some studies may foster empirically and philosophically driven symbiotic combination of normative and empirical research in AI Ethics. However, provided that the overall claims made in these studies are more descriptive than prescriptive, this research is also close to the social sciences. Therefore, it is settled that the **empirical work presented in this doctoral dissertation is located along the spectrum of empirical ethics and morally relevant social sciences**.

1.1.6. EMPIRICAL RESEARCH IN THE AI ETHICS CONTEXT

As mentioned earlier, AI Ethics faces three main challenges related to normative urgency, multi-purpose nature of modern AI, and multitude of stakeholders operating in the AI Ethics space. The premise of this doctoral dissertation is that empirical information is valuable for AI Ethics to address these challenges. Empirical research allows a systematization that is critical in the current context of normative urgency. Along these lines, recent empirical meta-studies analyzed the AI Ethics guidelines and brought much-needed clarity about the key principles and ethical issues in the messy landscape of guidelines [151, 152].

AI is now a multi-purpose technology affecting several societal domains. Therefore, empirical research is valuable to understand the perceptions of ethical risks and vulnerabilities associated with AI in different domains. In this context, healthcare is of particular interest. Several studies have surveyed medical practitioners on the impact of AI technologies in the medical profession [178–181] yet little attention was paid to ethical issues associated with the implementation of AI in healthcare. More empirical information is required to further the knowledge about AI's ethical risks and vulnerabilities in key societal domains.

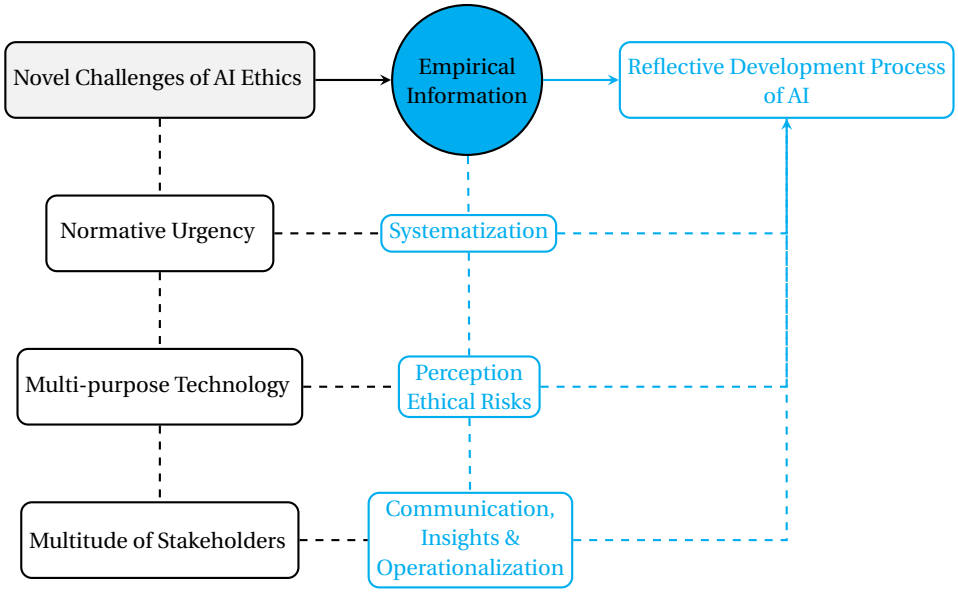
The multitude of stakeholders operating in the AI space requires a collaborative effort toward Ethical AI. For the normative communities, such collaboration entails (i) clear communication of the state-of-the-art normative work to the technology communities; (ii) incorporation of insights from the technology communities in normative work; and (iii) operationalization of theoretical concepts. For the technology communities, a collaborative effort entails a departure from "tick-the-box" Ethics and engagement in more nuanced normative critical thinking about AI.

The empirical work presented in this dissertation aims at assisting AI Ethics in overcoming the challenges outlined above and meeting its normative goals. When these challenges are mitigated, there is an improvement in the reflective development of AI Ethics, which allows the communities operating in the AI space to further engage in critical reasoning about AI.

Therefore, the ultimate ambition of this dissertation is to contribute to the reflective development of AI. By setting up such ambition, this research largely circumvents traditional criticism related to the use of empirical data in the normative domain (loss of normative edge; tendency to reject more general principles on the basis of relatively specific facts; entrenching prejudged ethical positions; lack of broad moral conclusions) [174, 175] but is vulnerable to other criticisms. There is no existing measurement of reflective development. Hence it is not possible to evaluate a potential improvement with scientific precision. This exploratory research may serve as a stepping stone for future endeavors that explore a reflective development measurement.

It has been reinstated that the **normative goal of AI Ethics is to provide guidance that allows society to capitalize on the innovation benefits of AI while minimizing its risks** and that **empirical information is valuable** to **overcome the current challenges** faced by AI Ethics and ultimately **improve the reflective development of AI**. At the core of this research is ***Ethical AI***. In rough terms, it is **AI through a morality lens**. Herein it is also a multi-dimensional concept that guides the structure of the dissertation.

Figure 1.6: Graphical representation of challenges of AI Ethics, empirical information, and reflective development.



1.1.7. THE DIFFERENT DIMENSIONS OF ETHICAL AI

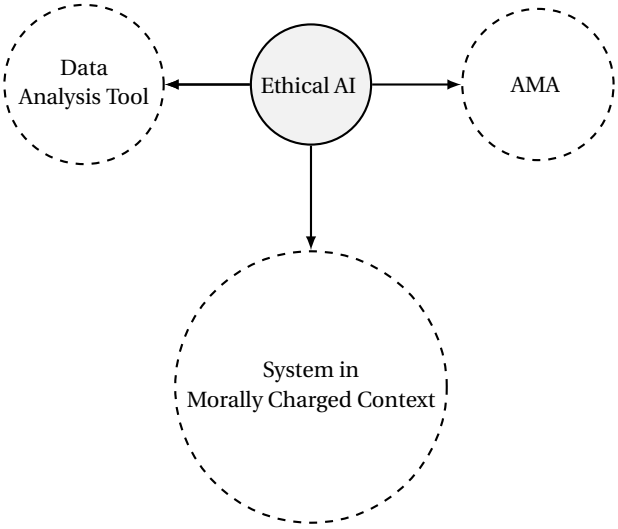
Through a morality lens AI may be regarded as: (i) **data analysis tool in morality**; (ii) **system operating in morally charged contexts**; or (iii) **artificial moral agent**. Stripped to its bare bones, AI is a **data analysis tool** in the current data-driven paradigm. Several breakthroughs have resulted from the use of learning algorithms in Science and Technology. A good example is AlphaFold, a deep learning system that predicts 3D structures of proteins based on the amino-acid chains [13]. In contrast, little research has explored the use of these algorithms in morality. This may be explained by the strained relationship between

empirical and normative sciences and also by the scarcity of morality data. A well-known exception is the MME. This empirical ethics study used unsupervised learning algorithms to analyze massive amounts of data on moral choices in traffic situations [134]. Despite the criticism received, the MME has shown the potential of using learning algorithms in moral data.

AI can also be regarded as a **system that operates in morally charged contexts**. As AI Systems become increasingly autonomous, their decisions and actions may have morally problematic implications. The archetypal case that has dominated popular and scientific literature is the AV. Traffic is a complex and dynamic environment where human lives are often in danger. Decisions made in this environment may carry a moral charge. The concerns about programming risk distribution decisions into AVs are well-known, but similar concerns apply to other morally charged contexts, such as healthcare or justice.

When regarded as an **artificial moral agent** (AMA), i.e. an artificial system with moral reasoning capabilities, AI is reconciled with its science fiction roots [128, 162, 182]. There is an array of existing and projected systems that qualify as AMAs. However, the AMA at the core of the Machine Ethics project is a system ranking high in autonomy and moral sensitivity of the sort featured in science fiction books and movies. This is perhaps the most polarizing dimension of *Ethical AI* given that part of the scientific community is dismissive of artificial morality [183]. The normative work developed in this dimension tends to be abstract, convoluted, and speculative. Proper systematization and operationalization are demanded from the Machine Ethics community to engage other communities on artificial morality.

Figure 1.7: Graphical representation of *Ethical AI*.



1.2. RESEARCH AIM, RESEARCH QUESTIONS, AND RESEARCH GOALS

RESEARCH AIM

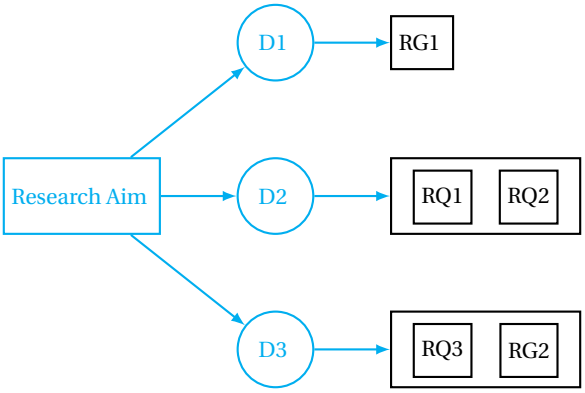
The aim of the dissertation is to **use an empirical approach to operationalize the different dimensions of *Ethical AI***.

RESEARCH QUESTIONS & RESEARCH GOALS

The research questions and goals formulated in this dissertation map onto the different dimensions of *Ethical AI*. The difference between a research question and a goal lies in the nature of the research contribution. The research questions aim at an empirical contribution for which there is a need for insights generated from a particular dataset. The research goals aim at a methodological contribution for which empirical data is used solely for illustrative purposes. The first research goal (RG1) relates to the first dimension, which features AI as a data analysis tool in morality (D1). The first and second research questions (RQ1 & RQ2) relate to the second dimension, which features AI as a system in morally relevant contexts (D2). The third research question (RQ3) and the second research goal (RG2) relate to the third dimension, which features AI as an artificial moral agent (D3). The research questions and goals are outlined below.

- RG1: To illustrate how AI-techniques can be used to investigate the relationship between morality and behavior.
- RQ1: What are the ethical issues in focus by the autonomous vehicles industry?
- RQ2: What are the perspectives of medical practitioners about Health AI?
- RQ3: What are the perspectives of AI Ethics scholars about Artificial Moral Agents?
- RG2: To provide an empirical method which is able to equip AI with moral uncertainty.

Figure 1.8: Graphical representation of research aim, dimensions of Ethical AI, research questions, and research goals.



1.3. RESEARCH METHODOLOGY

The research featured in this doctoral dissertation is located along the spectrum of empirical ethics and morally relevant empirical social sciences. The methodology is, therefore, markedly empirical. Both qualitative and quantitative methods from the social sciences toolkit (statistical model estimation, systematic review, survey) were used to derive empirical information within the domain of AI Ethics.

In order to address RG1, unsupervised learning algorithms were used to investigate clusters of moral foundations in a morally charged legal dataset. For RQ1, a systematic literature and industry review was conducted in order to assess the awareness and engagement of AV companies on the ethical issues of AVS. To tackle RQ3 and RQ4, q-methodological surveys were conducted to identify relevant perspectives about Health AI and artificial morality. For RG2, a latent class choice model was used to operationalize moral uncertainty (Table 1.1).

Five datasets were used in this dissertation. As indicated earlier, the empirical data used to address the research questions and goals served different purposes. The data used to address the research questions generated insights suited to answer those particular questions, whereas the data used to address the research goals illustrate the methodological contributions. The dataset used to illustrate RG1 features information about gender-based violence legal cases. The dataset used to address RQ1 includes morally relevant statements retrieved from industry reports of autonomous driving companies. The dataset used to address RQ2 features opinions of medical practitioners about Health AI collected in a q-methodological study. The dataset used to address RQ3 features opinions of AI Ethics scholars about artificial morality collected in a q-methodological study. The dataset used to illustrate RG2 features choice data about moral trade-offs in traffic [5]. All datasets are publicly available.

The empirical approach taken in this dissertation is mainly an exploratory research effort, which is deemed adequate to address the general research questions formulated earlier. The studies aim to provide insights, improve the knowledge, and further operationalize the different dimensions of Ethical AI. Accordingly, this research is expected to be a stepping stone for future research in AI Ethics.

Table 1.1: Research methods and data.

RG	RQ	Method	Dataset
RG1		Quantitative: clustering	Gender-based violence legal cases
	RQ1	Qualitative: systematic review	Ethical issues in AV industry reports
	RQ2	Mixed: q-methodological survey	Perspectives about Health AI
	RQ3	Mixed: q-methodological survey	Perspectives about AMAs
RG2		Quantitative: latent class choice model	Taboo-trade off choice experiment

[5] The taboo-trade-off-dataset was used as a secondary source of data from [184].

1.4. STRUCTURE OF DOCTORAL DISSERTATION

The doctoral dissertation features this introductory chapter, the main body, and a conclusion. The main body of the dissertation consists of five chapters originally conceived as self-standing articles (four articles are published, and one is currently in the peer-review process). Each chapter relates to a dimension of *Ethical AI* and addresses the research questions and goals outlined above (Figure 1.8). The second chapter relates to D1 and addresses RG1, the third and fourth chapters relate to D2 and address RQ1 and RQ2, and the fifth and sixth chapters relate to D2 and address RQ3 and RG2.

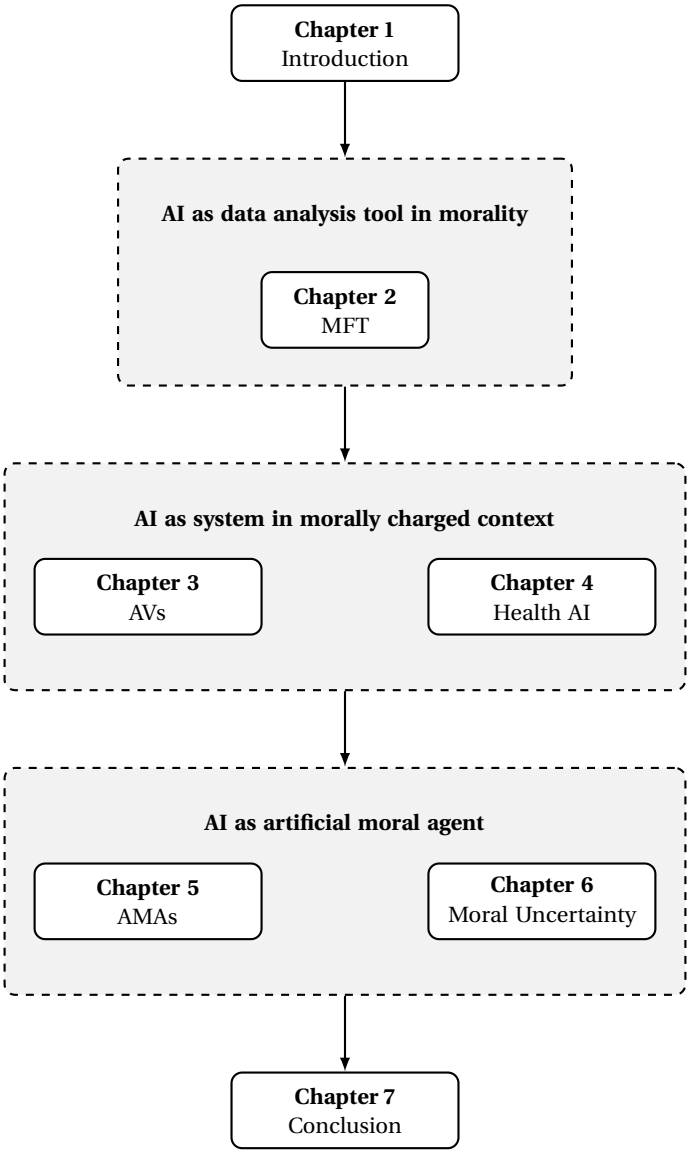
In **chapter 2** AI is conceptualized as a data analysis tool that is used to gain empirical insights into morality. Learning algorithms have been used in several knowledge domains, yet little research has explored these algorithms in morality. To illustrate how these algorithms can be used the moral domain, a unsupervised learning algorithm is used to explore the moral rhetoric in legal narratives. The moral foundations, as posited by the Moral Foundations Theory, are identified and clustered within a corpora of morally charged legal cases about gender-based violence. A contribution is made to the field of morality, thus making a case for using AI in this domain.

In **chapters 3 and 4** AI is conceptualized as a system that operates in morally charged contexts. In **chapter 3** this system is an AV. Much attention has been paid to the ethics of autonomous driving, yet little information is available about the ethical issues in focus by the AV industry. A systematic review was conducted to compare and contrast the normative narratives in scientific and industry communities. The analysis of reports from a wide range of technology and manufacturing companies at the forefront of autonomous driving technology allows important insights into ethics within the AV industry. In **chapter 4** the AI system operates in the healthcare setting (Health AI). Provided that AI has the potential to threaten core values in Medicine, it is important to gain insight into the reasoning patterns and moral opinions about Health AI from those involved in the medical practice. A q-methodological survey is conducted to investigate the perspectives of medical doctors about key ethical issues associated with Health AI. The different perspectives about Health AI provide valuable insights on the ethical issues that should be operationalized and accounted for in the design and development of these technologies. Both studies featured in these chapters aim to further the knowledge about the risks and vulnerabilities of AI in key societal domains (Transportation and Healthcare). These insights may be incorporated in future normative work about AI.

In **chapters 5 and 6** AI is conceptualized as an AMA. The development of AMAs is central to the Machine Ethics project, and it is quite controversial. In **chapter 5** a q-methodological survey is conducted among AI Ethics scholars in order to bring clarity and coherence to the complex topic of artificial morality. This study aims at improving the communication of state-of-the-art normative work to other communities operating in the AI space. In **chapter 6** an empirical method to equip AI with moral uncertainty is proposed. A theoretical metanormative framework for equipping AI with moral uncertainty is operationalized through a latent class choice model. This operationalization provides an empirical illustration of moral uncertainty.

The conclusion (**chapter 7**) of the dissertation features a summary of the main findings reported in each chapter, scientific contributions of research, policy reflections, and a final reflection from the author.

Figure 1.9: Graphical representation of structure of doctoral dissertation.



I

AI as Data Analysis Tool in Morality

2

MORAL FOUNDATIONS IN LEGAL NARRATIVES: A CASE STUDY ON GENDER-BASED VIOLENCE

ABSTRACT

The MFT is a theoretical model of intuitive ethics that builds on the assumption that a small number of moral foundations related to different adaptive challenges emerge as intuitive reactions to social stimuli. Recent MFT research has explored the relationship between moral foundations and human behavior. However, these studies rely on self-scoring and social experiments, which have well-documented limitations. In this study, we expanded this line of research by investigating the relationship between moral foundations in gender-based violence legal narratives and corresponding legal outcomes. We used a Latent Class Cluster Analysis model to identify the complex patterns in the data and reduce such patterns to a distinct number of clusters. Three main clusters unravel latent relations between the foundations mapped in the legal narratives and corresponding legal outcomes: (i) Affirmative with suspended prison time; (ii) Mixed outcomes but no prison time; and (iii) Affirmative with much prison time and compensation. We established that there is an association between moral foundations in gender-based violence narratives and corresponding legal outcomes, but further research is needed for a better understanding of the causal mechanisms underlying this association.

2.1. INTRODUCTION

Among the investigations that have looked into the evolutionary roots of human morality, the Moral Foundations Theory (MFT) is perhaps the most popular [185–190]. This theoretical model of intuitive Ethics builds on the assumption that moral intuitions derive from innate psychological mechanisms that co-evolved with cultural institutions [189, 191]. Those mechanisms are the moral foundations upon which individuals and cultures construct their moralities [192].

According to the MFT, a small number of moral foundations, linked to different adaptive challenges related to care, fairness, loyalty, authority, and sanctity, emerge as intuitive reactions to social stimuli [193]. These foundations are the building blocks of a rudimentary

moral function that grows with increasing sophistication through the acquisition and use of a wide variety of moral concepts, perceptions, beliefs, and judgments [193]. The MFT posits that the moral foundations account for the variety of moral judgments underlying individual and cross-cultural differences in moral reasoning [193, 194].

Great progress has been made in testing the MFT [189]. Validated measures grounded on the MFT, such as the Moral Foundations Dictionary (MFD) [1] and the Moral Foundations Questionnaire (MFQ) [2], were developed to explore the moral domain systematically. These measures have opened avenues for MFT empirical research.

This line of empirical research has focused mainly on exploring the relation between moral foundations and socio-political beliefs [192, 195, 195, 196] on controversial topics such as stem cells [197], abortion [198], or immigration [199]. More recently, some studies have also explored the relationship between moral foundations and behavior [200, 201]. Assuming that the moral foundations are rooted in evolution entails that they should bear some influence on human behavior. These investigations are important to further understand the influence of the moral foundations on human moral behavior.

Empirical studies that explore this line of research measure moral intuitions of participants through the MFQ and evaluate morally relevant behavior through self-reports or measurements in social experiments.

The predictive power of the moral foundations is not consistent across studies [200–204], but it is clear that the foundations are important variables for a nuanced understanding of human behavior. Relevant findings reported in the literature focused mainly on the contrast between individualizing and binding foundations with respect to moral behavior. It was reported that individualizing foundations are associated with a leniency effect on sentencing decisions in a mock juror trial [203], charity donations [201], and positive feelings towards the Green Party [204]. Despite the relevance of these studies, they have important limitations related to self-reporting and social experiments, namely poor congruence between endorsed moral foundations and action as well as adjustment of behavior to social desirable biases.

This study aims to expand this line of research by exploring the relationship between moral foundations in narratives, measured through the MFD, and morally relevant behavior. Rather than relying on self-reports or social experiments, we look at moral rhetoric in a text corpus and the behavioral outcomes related to those narratives. This approach is expected to provide further insights into morality while mitigating well-documented limitations of self-reporting and social experiments.

The narratives investigated in this study relate to legal cases. Several MFT studies have mapped the moral rhetoric in different outlets, such as church sermons [192], social media [199, 205], or newspapers [197], but legal narratives have remained largely unexplored. Our particular case study is about gender-based violence legal cases. These narratives are convenient for exploring whether moral foundations are associated with moral outcomes. The gender-based violence legal cases are morally charged, and the corresponding legal

[1] The MFD provides virtue and vice vocabulary associated with each foundation thus allowing the analysis of moral rhetoric in text [192].

[2] The MFQ is a scale that measures the degree to which individual moral beliefs and concerns rely upon the moral foundations thus gauging differences in the range of concerns that people consider morally relevant [191].

decisions are morally relevant behavioral outcomes.

The remainder of this article is organized as follows: First, we contextualize the moral foundations within gender-based violence (Section 2.2). Subsequently, we briefly characterize our case study, which is the social problem of gender-based violence in Portugal (section 2.3). We proceed with describing the methodology used in this study (section 2.4). Later the results are presented (section 2.5) and discussed (section 2.6). Finally, we outline the main conclusions of this study (section 2.7).

2.2. THE MORAL FOUNDATIONS CONTEXTUALIZED IN GENDER-BASED VIOLENCE

The MFT was originally developed by Haidt and Graham to describe moral differences across cultures [194]. They looked for matches of moral concerns found widely across cultures for which there were plausible and published evolutionary explanations of related psychological mechanisms. They identified five main concerns which they hypothesized to be the foundations of intuitive morality [192, 194]. The foundations identified by Haidt and Graham are innate moral modules linked to different adaptive challenges related to care, fairness, loyalty, authority, and sanctity that emerge as intuitive reactions to social stimuli [193].

The moral foundations are often divided into individualizing foundations (care and fairness), which focus on the protection of individuals, and binding foundations (loyalty, authority, sanctity), which focus on the protection of the needs and well-being of collectives [191, 206].

These foundations are presented as general conceptualizations. Therefore, in social and moral research, there is a need to interpret them in light of particular social phenomena. In this study, we contextualize the moral foundations within gender-based violence in order to understand their meaning and implications in the legal cases.

Care produces intuitions that protection, prevention of harm, and caring for others are morally good [189]. Care intuitions are implicated in judgments about gender-based violence. This foundation underlies narratives of protecting vulnerable groups, such as women and children, in gender-based violence cases. However, when accounting for situational variability (e.g., non-stereotypical offenders or victims) in gender-based violence cases, it is less clear how and under which conditions this foundation is activated [207].

Fairness is related to equality, reciprocity, and justice in cooperative acts [189]. This foundation is particularly relevant to understand variations in punitive responses to gender-based violence. It encompasses an element of proportionality, which implies an adjustment between the harmfulness of the offending behavior and the responsibility of the offender in the punitive responses to gender-based violence. For example, it may be a moderating factor in punitive responses to first-time juvenile offenders compared to adult repeat offenders. Fairness also encompasses an element of social equality. In line with feminist views that violence against women is a manifestation of gender oppression, it may also imply higher levels of punitive attitudes in cases that are perceived as strongly misogynistic [207].

Loyalty produces intuitions that sacrifice for a group, patriotism, and faithfulness are

morally good [189]. This foundation is related to the overestimation of danger instilled by strangers and the underestimation of danger within trusted family or community circles. Therefore it may explain poor reporting of violence and abuse within social or familial groups in the interest of maintaining family or community unity [207]. When such abuse is reported, the breach of loyalty expected from those groups becomes a focal point of legal narratives and sentencing. Moreover, given that this binding foundation focus on the integrity of the collective, it underlies punitive responses that create social distance between offenders and the community (e.g., incarceration) rather than more progressive and rehabilitation oriented policies that aim to successfully reintegrate offenders in the community [207].

Authority is related to respect and obedience to authority, social traditions, and hierarchies [189]. This foundation places great value on respect for the rule of law and established social hierarchies. Therefore, offending is considered immoral strictly because it violates the rules that society has established. Adherence to this foundation is expected to be associated with more pronounced punitive attitudes. However, because this foundation is also associated with respect for social hierarchies, when coupled with conservative views of males as dominant over females in the social hierarchy, it may lead to lenient punitive responses to violence occurring in family [207].

Sanctity produces intuitions that bodily purity, however defined by different cultures, is morally good [189]. It refers to moral intuitions stemming from our innate aversion to phenomena that are deemed to be unnatural. Gender-based violence is a complex social problem which encompasses several forms of abuse. Cases involving sexual abuse prominently intersect with notions of purity. Adherence to this foundation may lead to more severe punitive responses in cases involving sexual violence [207].

Our case study is about gender-based violence in Portugal. By investigating these legal narratives we expect to make a contribution to the MFT research on the behavioral implications of moral foundations. A background on the case study explored in this research is provided below.

2.3. GENDER-BASED VIOLENCE IN PORTUGAL

Gender-based violence is a prevalent social problem in Portugal. It encompasses a multitude of violent behaviors, including physical, sexual, psychological, and economic harm that affects mainly women and girls. The most common type of gender-based violence in Portugal is domestic violence. It is consistently one of the most reported crimes to the police force and the leading cause of women's homicide in Portugal[3] [208].

Domestic violence has received some attention from the Portuguese scientific community. Studies have looked into socio-demographics [209][4], psycho-sociological [210–213], forensic [208, 214, 215], and criminology aspects of domestic violence in Portugal [209, 216, 217]. The findings reported in these studies are valuable insights about this complex social problem in Portugal.

[3] Annual Report of Internal Safety 2020 and Annual Monitoring Report on Domestic Violence issued by the Portuguese Ministry of Internal Administration 2016-2022.

[4] Annual Monitoring Report on Domestic Violence issued by the Portuguese Ministry of Internal Administration 2016-2022.

The psycho-sociological studies indicate that child and partner abuse has for a long time been a significant problem in Portugal [210, 211]. It affects mainly women and girls but also particularly vulnerable populations, such as elders living with disabilities [212, 214]. A recent study that looked at the evolution of attitudes and behaviors of adolescents concerning intimate partner violence reported that, although traditional beliefs on marital violence are rejected more now than in the past, dating violence reports have not decreased [213].

A retrospective study of forensic data provided important insights into fatal outcomes of domestic violence. This study looked at forensic data concerning female intimate partner homicide victims (years 2005 - 2007) and reported that most victims lost their lives to gunshot trauma caused by shotguns and rifles. It was hypothesized that access to those specific firearms could be explained by the fact that sport hunting is regularly practiced in Portugal [208].

Some studies have also looked into criminology aspects of domestic violence [209, 216, 217]. A retrospective analysis of the complaints registered in the database of one of the Portuguese police forces (years 2010-2013) showed that psychological and physical abuse are the most frequent forms of violence [216]. It was also indicated that violence might decrease after the alleged offender contacts with the criminal justice system for the first time [216]. This finding is aligned with other studies that showed that prosecution has a more discouraging effect on violent behaviors than conviction [218, 219]. Another recent study looked at a sample of 240 domestic violence legal decisions (years 2015-2019) to investigate potential discrimination [209]. The study reported no evidence of discrimination with respect to age, nationality, or education of offenders in such legal decisions. However, conviction rates were found to be higher for female judges making an individual legal decision [209].

A controversial legal decision that cited the Bible and a 1884 Penal Code to frame a case of domestic violence [5], prompted a recent interest in the analysis of legal narratives. One study analyzed the controversial decision and reported that the judge formulated a reasoning apparently based on a generic content of common experience only to give voice to a highly subjective judgment. Such reasoning enforced social constructs about gender-based violence at odds with dominant social conceptions. By representing the woman, who was a victim of domestic violence, as the responsible agent, the judge normalized the abuse and made ample room for a lenient sentence [217].

This study also explores the narratives of domestic violence cases in Portugal. We aim to investigate the relationship between moral rhetoric, as mapped by the moral foundations, and corresponding legal outcomes. We expect to contribute to MFT research while also providing important insights into gender-based violence in Portugal.

2.4. METHODOLOGY

Our approach to realize the main ambition of this study, which we recall is to explore the relationship between moral rhetoric in gender-based violence legal cases and corresponding legal outcomes, consisted of three steps: (i) data collection; (ii) mapping the moral rhetoric; and (iii) analysis.

Initially, we compiled a corpus of ($N = 655$) domestic violence cases decided in Portuguese

[5] ECLI:PT:TRP:2017:355.15.2GAFLG.P1.

courts. Subsequently, we mapped the moral rhetoric in the legal narratives through the MFD. Finally, we created a dataset featuring relevant information about the cases, moral foundations, and legal outcomes and proceeded with the statistical analysis.

We used a Latent Class Cluster Analysis (LC) model to unravel the relationship between moral rhetoric and legal outcomes. This model identifies the complex patterns among a set of pre-defined indicators related to legal outcomes and reduces such patterns to a distinct number of clusters. Covariates are also included in the model to provide further insights into cluster membership.

2.4.1. DATA COLLECTION & DATASET

The dataset used in this study consists of domestic violence cases decided by Portuguese courts in the years 2002 to 2020. The legal cases were retrieved from the Portuguese Jurisprudence Research Portal of the Superior Council of the Judiciary [6]. This portal features a randomly assigned sample of legal decisions of intermediate level courts of appeal (Tribunal da Relação de Lisboa, Tribunal da Relação do Porto, Tribunal da Relação de Coimbra, Tribunal da Relação de Guimarães, Tribunal da Relação de Évora) as well as the Supreme Court of Justice in Portugal. Unfortunately, at the time of this research the legal cases decided by the lower level Comarca courts were not available online.

For the identification of the domestic violence cases we used key words "violência doméstica" (domestic violence) and "maus tratos entre cônjuges" (spousal mistreatment) in the Jurisprudence Research Portal. The timeline for the search (2002-2020) reflects the current legal paradigm of domestic violence as a public crime in Portugal. This means that a formal complaint is not required for the District Attorney to initiate a criminal investigation. The cases included in this study feature a decision about domestic violence, mistreatment, or aggression but often include other offenses.

A total of 655 cases duly identified through the European Case Law Identifier (ECLI) number were included in this study. The dataset features twenty-seven variables, which were populated with information retrieved from the legal cases [7]. Although not all the data collected was used in this study, future research may further explore this dataset, as it will be made publicly available.

The distribution of legal cases within the timeline 2002-2020 shows that there is a limited number of cases available in the Jurisprudence Research Portal in particular years (e.g. 2020) (Table 2.1). On the contrary, the distribution of cases with respect to the different courts is quite balanced, except for the Tribunal da Relação de Guimarães, which is featured less than the other courts in this dataset (Table 2.2). In the vast majority of cases, the person accused of a gender-based violence offense is male and the outcome is a conviction (Tables 2.3 and 2.4)[8].

[6] https://jurisprudencia.csm.org.pt/.

[7] The variables featured in the dataset are: year, court, votation of judges, appeal, gender of accused, gender of appellant, outcome, offenses, prison time, suspension of prison time, accessory penalty of restraining order, accessory penalty of civic training, replacement penalty of community work, fine, compensation, charity donations, gender of judges, moral foundation care virtue, care vice, fairness virtue, fairness vice, loyalty virtue, loyalty vice, authority virtue, authority vice, sanctity virtue, sanctity vice.

[8] Other legal outcomes include: accused, not accused, new legal decision, new trial, habeas corpus accepted, habeas corpus denied, coercive measure revoked, coercive measure denied, coercive measure applied, safety measure, new investigation, additional pre-trial investigation, additional pre-trial investigation denied, victim

Table 2.1: Number of legal cases featured in dataset in the timeline 2002-2020.

Year	Number of Cases
2002	1
2003	5
2004	6
2005	6
2006	5
2007	6
2008	20
2009	10
2010	29
2011	39
2012	47
2013	63
2014	42
2015	63
2016	81
2017	69
2018	85
2019	64
2020	14
Total	655

This study is anchored in domestic violence. However, it is acknowledged that it is not possible to isolate this particular offense from a broader complex social and criminal context. Several other offenses emerge along with domestic violence, which are also featured in this study, such as physical aggression, threat, or insults. By including these offenses, this study captures the nuances of this social problem and circumvents issues of misinterpretation (e.g., a case of aggression towards spouse that is framed as physical aggression rather than domestic violence), thus providing a rich account of gender-based violence in Portuguese jurisprudence.

To validate the dataset, three different Portuguese-speaking assistants were asked to check the accuracy of the information featured in the dataset with reference to the legal cases.

testimony, provisional suspension of process denied, pre-trial investigation secrecy, pre-trial investigation secrecy denied, probation denied, suspension of prison time, and suspension of prison time revoked.

Table 2.2: Number of legal cases per court featured in dataset.

Court	Number of Cases
Supreme Court	103
Tribunal da Relação de Guimarães	58
Tribunal da Relação do Porto	133
Tribunal da Relação de Coimbra	104
Tribunal da Relação de Lisboa	108
Tribunal da Relação de Évora	149

Table 2.3: Gender of person accused of gender-based violence offense in the cases featured in the dataset.

Gender of Accused	Number of Cases
Male	611
Female	30
Male & Female	14

Table 2.4: Legal decisions in gender-based violence cases featured in the dataset.

Legal Decision	Number of Cases
Conviction	445
Acquittal	44
Other	166

2.4.2. MAPPING THE MORAL RHETORIC

In order to map the moral rhetoric in the domestic violence legal cases, we relied on the MFD 2.0, which features 295 words and word stems related to each vice/virtue dimension of the moral foundations [192, 220].

Provided that, to the best of our knowledge, a Portuguese version of the MFD 2.0 was not available, we first proceeded with the translation of the dictionary to the Portuguese language [9]. Subsequently, we used a normalized term frequency approach to measure the moral foundations in legal cases. The terms associated with the moral foundations (as per the Portuguese MFD 2.0) were divided by the total number of terms in each legal case. Accordingly, the weight of each moral foundation in a legal case varies between 0 and 1 (Table 2.5).

[9] The Portuguese MFD 2.0 dictionary is made available online.

Table 2.5: Mean and standard deviation of moral foundations in gender-based violence cases featured in the dataset.

Moral Foundation	Mean	Standard Deviation
Care Virtue	.0119	.0040
Care Vice	.0145	.0048
Fairness Virtue	.0076	.0032
Fairness Vice	.0009	.0006
Loyalty Virtue	.0043	.0019
Loyalty Vice	.0004	.0003
Authority Virtue	.0122	.0040
Authority Vice	.0014	.0015
Sanctity Virtue	.0055	.0025
Sanctity Vice	.0046	.0028

In order to validate the term frequency approach, we investigated the distribution of moral foundations in different types of legal cases by comparing the dataset used in this study with two different control datasets. Each control dataset featured twenty legal cases retrieved from the Jurisprudence Research Portal. One dataset featured nonviolent crimes (driving without a license, driving under the influence, copyright infringement), and the other featured violent crimes (homicide, sexual abuse, incitement to prostitution). Upon comparing the distribution of moral foundations in the different datasets, we confirmed a variation in the distributions (for instance, the authority/vice foundation is much higher in the dataset featuring violent crimes), thus providing a first sense of face validity of our study.

2.4.3. LATENT CLASS CLUSTER ANALYSIS

OVERVIEW AND MATHEMATICAL FORMULATIONS

Latent Class Cluster Analysis (LCA) is a statistical technique that uncovers hidden clusters in data. It is a special kind of finite mixture model, also known as unsupervised learning models, in which the statistical distribution is modeled by a mixture (or weighted sum) of other distributions. A LC model identifies complex patterns among the set of (continuous, nominal, or ordinal) indicators [10] and reduces those patterns to a distinct number of clusters [222]. In a LC model the independent variables x are both the indicators and covariates, whereas the dependent variable Y is the latent class variable. The clustering is based on the differences in indicators [223] and covariates are included in the LC model to further explain cluster membership.

There are two key assumptions in LCA. The first assumption (mixture of c classes) is that the joint probability distribution $P(y_i...y_j)$ is a mixture of j class-specific distributions. The second assumption (local independence) means that indicators $y_i...y_j$ are independent within each latent class $c_1, c_2, c_3...c_n$.

[10] The term Latent Profile Analysis is preferred when indicators are continuous [221].

The LCA approach features two sets of parameters. The first is the set of inclusion probabilities (cluster membership probabilities), i.e., the probability that any random case in a population will be included in any LC. The second is the conditional probability that, given a specific class, an indicator takes a certain value (conditional probability).

$$P(\mathbf{X}_n) = \sum_{c=1}^{C} \left(P(Y = c) \cdot \prod_{i=1}^{I} P(X_{ni} \mid Y = c) \right) \tag{2.1}$$

where $\sum_{c=1}^{C} P(Y = c)$ is the cluster membership probability for a class c among C possible classes, and $\prod_{i=1}^{I} P(X_{ni} \mid Y = c)$ is the product of the conditional probability of a pattern of I indicators for N cases, i.e X_{ni} is the indicator i of case n.

The class membership probability $P(Y = c)$ is provided by a softmax function, which turns a vector of C real values into a vector of C real values that sum to 1 and so it may be interpreted as a probability.

$$P(Y = c) = \frac{\exp(\gamma_{0c})}{\sum_{c=1}^{C} \exp(\gamma_{0c})} \tag{2.2}$$

In this study we defined the LC model with covariates. Both covariates and indicators are independent variables but for purposes of differentiation we write $\mathbf{Z}_k$ as the vector of covariate k for N cases $\mathbf{Z}_k = \{z_{1k} ... z_{Nk}\}^T$ and $\mathbf{X}_i$ as the vector of indicators i for the N cases $\mathbf{X}_i = \{x_{1i} ... x_{Ni}\}^T$. The covariate effects are therefore estimated simultaneously with the indicators defining the class-specific item distributions. The conditional probability $P(\mathbf{X}_n \mid \mathbf{Z}_n)$ is defined as:

$$P(\mathbf{X}_n \mid \mathbf{Z}_n) = \sum_{c=1}^{C} P(Y = c \mid \mathbf{Z}_n) \cdot P(\mathbf{X}_n \mid Y = c) \tag{2.3}$$

The conditional probability $P(Y = c \mid \mathbf{Z}_n)$ is provided by a softmax function.

$$P(Y = c \mid \mathbf{Z}_n) = \frac{\exp(\gamma_{0c} + \sum_{k=1}^{K} \gamma_{kc} \cdot z_{nk})}{\sum_{s=1}^{C} \exp(\gamma_{0c} + \sum_{k=1}^{K} \gamma_{ks} \cdot z_{nk})} \tag{2.4}$$

The conditional probability $P(\mathbf{X}_n \mid Y = c)$ for I_c categorical indicators with response values V is provided by:

$$P(\mathbf{X}_n \mid Y = c) = \prod_{i=1}^{I_c} P(X_{ni} \mid Y = c) = \prod_{k=i}^{K} \prod_{v=i}^{V_k} \theta_{kcv}^{I(Y_{ik}=v)} \tag{2.5}$$

The conditional probability $P(\mathbf{X}_n \mid Y = c)$ for $I - I_c$ continuous indicators is provided by:

$$p(\mathbf{X}_n) = \sum_{c=1}^{C} P(Y = c) \cdot p(\mathbf{X}_n | Y = c) \tag{2.6}$$

where $p(\mathbf{X}_n|Y=c)$ is by default a local independence model of the following form:

$$p(\mathbf{X}_n \mid Y=c) = \prod_{i=I_c+1}^{I} \frac{1}{\sqrt{2\pi\sigma_{i,Y}^2}} \exp\left[-\frac{1}{2\sigma_{i,Y}^2}(X_{nc}-\mu_{c,Y})^2\right] \tag{2.7}$$

CHARACTERIZATION OF THE LC MODEL

The LC model used in this study features five indicators and four (active) covariates. Given that some of these variables are categorical, we relied on effect-coding to include them in the LC model. A graphical representation of the model is shown in Figure 2.1.

The indicators relate to the legal outcomes of gender-based violence cases. The indicator *outcome* is a categorical variable featuring three different categories concerning possible outcomes of legal cases: (i) *affirmative of a gender-based violence situation*, i.e., suspect of offense is accused, convicted, or there was a need for coercive or safety measure; (ii) *not affirmative of a gender-based violence situation*, i.e., suspect of offense is not accused, not convicted, or there was no need for a coercive or safety measure; and (iii) *neutral* i.e., decisions related to pre-trial investigation secrecy matters or decisions that did not have an outcome *per se* as there was a need for additional investigation, new legal decision, or re-trial.

The remaining indicators, which are prison time (months), suspension time (months), fine (EUR), and compensation (EUR), are continuous variables.

The covariates were included in the LC model to further explain cluster membership. The covariate *judges* is a categorical variable featuring eight different categories that capture the most common combinations of judges with respect to a binary account of gender, where *f* stands for female and *m* stands for male ((i) f; (ii) m; (iii) ff; (iv) mm; (v) fm; (vi) mf; (vii) mmm), but also residual combinations ((viii) other).

The covariate *court* is a categorical variable featuring six different categories ((i) Supreme Court of Justice; (ii) Tribunal Relação Guimarães; (iii) Tribunal Porto; (iv) Tribunal Relação Coimbra; (v) Tribunal Relação Lisboa; and (vi) Tribunal Relação Évora).

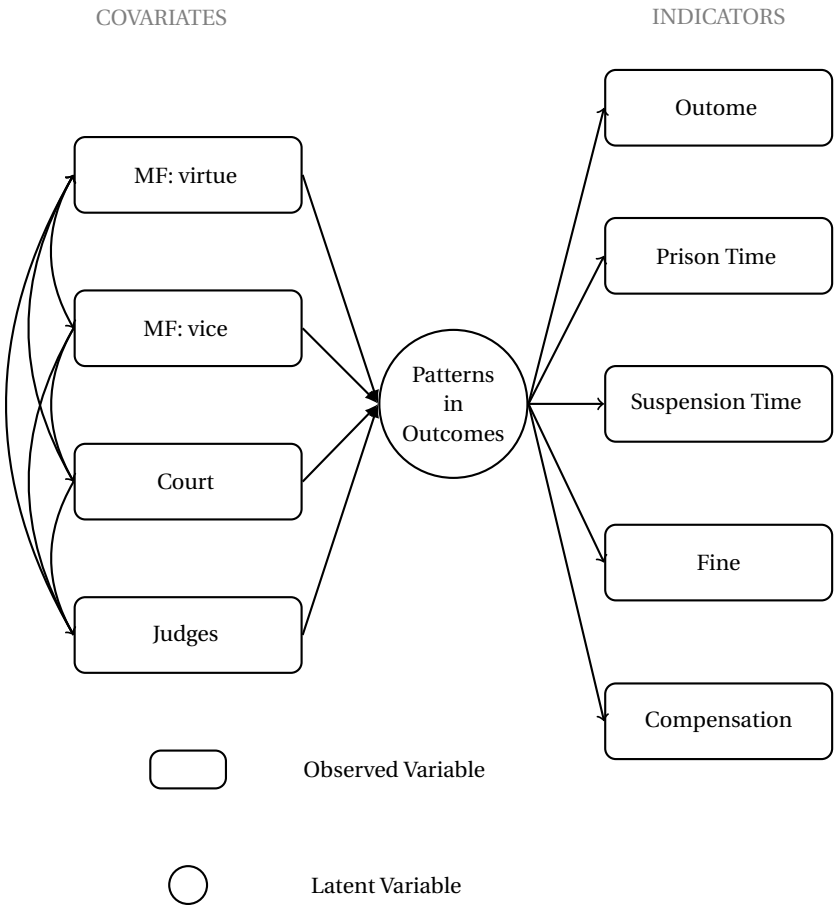
The covariate *moral foundations* features five virtue dimensions of the moral foundations ((i) care virtue; (ii) fairness virtue; (iii) loyalty virtue; (iv) authority virtue; and (v) purity virtue) and five vice dimensions of the moral foundations ((i) care vice; (ii) fairness vice; (iii) loyalty vice; (iv) authority vice; and (v) purity vice). These dimensions are continuous variables taking values between 0 and 1 as described in the sub-section 4.2.

Additional covariates related to *offenses* were also included in the study but kept inactive: domestic violence; mistreatment; illegal possession of a weapon; physical aggression; threat; sexual aggression; and homicide). Inactive covariates are not part of the specified model and have no influence on the model parameter estimates but may be used as explanatory variables in a post-processing stage. We decided to keep the variables mentioned above inactive because it is unclear whether they are indicators or covariates. Still, they provide important insights into the offenses associated with the different estimated clusters.

MODEL ESTIMATION

The LC model is estimated based on maximum likelihood i.e assuming parameter values that maximize the likelihood of observing the data. The estimates of the parameters are

Figure 2.1: Graphical representation of LC model.



obtained by maximizing the log-likelihood function (LL) based on $P(\mathbf{X}_n \mid \mathbf{Z}_n)$:

$$LL = \sum_{n=1}^{N} \log P(\mathbf{X}_n \mid \mathbf{Z}_n) = \sum_{n=1}^{N} \log \sum_{c=1}^{C} P(Y = c \mid \mathbf{Z}_n) \cdot P(\mathbf{X}_n \mid Y = c) \tag{2.8}$$

We estimated five models (1-5 classes) in Latent Gold. To assess the model fit we used information criteria that balances both model fit and complexity. The basic formulation for information criteria is:

$$IC = -2LL + wk \tag{2.9}$$

where w is a constant (weight) and k is the number of parameters.

In this study we used the Bayesian Information Criteria (BIC), which is a well-established measure of the goodness of fit of a statistical model in which $w = ln(s)$ and s is the sample size with lower BIC indicating better model fit [224]. However, in our study the BIC values consistently decrease as C increases, which makes model selection based on BIC difficult. By computing the LL of the models, we observe that there is no substantial increase in the relative fit of the model beyond the three-class threshold (Table 2.6 and Figure 2.2). Moreover, given that it is also the more interpretable model, a decision was made to use this model featuring three clusters.

Table 2.6: Model fit of LC models.

Classes	**Parameters (*N*)**	**BIC**	**LL**
1	10	39464.9814	-19700.0675
2	43	33040.6894	-16380.9250
3	76	30410.8962	-14959.0319
4	109	28699.9241	-13996.5495
5	142	27980.4875	-13529.8346

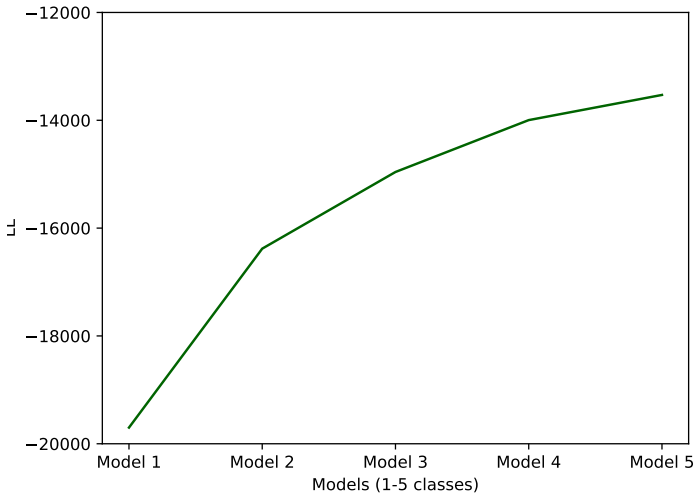


Figure 2.2: Log-Likelihood of Models (1-5 classes).

2.5. RESULTS

In this section, we analyze the LC model featuring three clusters. First, we provide information about the indicators and covariates of the model. We also provide information about the most common offenses in each cluster based on the cluster probabilities assigned to each legal case (dataset). Subsequently, we proceed to characterize each cluster. Finally, we report the distribution of individualizing and binding foundations in gender-based violence legal cases corpus.

2.5.1. OVERVIEW OF MODEL: INDICATORS AND COVARIATES IN THREE-CLUSTER MODEL

Three latent clusters are identified in the data (Table 2.7). The first cluster is labeled ***Affirmative with suspended prison time*** and has the largest share in the sample (0.45); the second cluster is labeled ***Mixed outcomes but no prison time*** and has a share of 0.40; and the third cluster is labeled ***Affirmative with much prison time and compensation*** and has the smallest share in the sample (0.15). The parameters of the indicators and statistically significant covariates as well as the cluster probability of covariate court and inactive covariates related to offenses are featured below (Tables 2.8, 2.9, 2.10, 2.11, 2.12).

Table 2.7: Size of clusters.

Cluster	Size
Cluster 1	.45
Cluster 2	.40
Cluster 3	.15

Table 2.8: Parameters of indicators in clusters (1-3).

Indicators	Cluster 1	Cluster 2	Cluster 3
Affirmative of gender-based violence	.993	.417	.999
Not Affirmative of gender-based violence	.004	.207	.000
Neutral	.004	.376	.000
Prison time in months (mean)	36	.000	131
Suspension time in months (mean)	20	.000	8
Fine in EUR (mean)	.000	152.142	507.866
Compensation in EUR (mean)	1826.566	47.764	25850.060

Table 2.9: Cluster probabilities of covariate Court.

Court	Cluster1	Cluster2	Cluster3
Supremo Tribunal de Justiça	.40	.18	.42
Tribunal Relação Guimarães	.64	.31	.05
Tribunal Relação Porto	.40	.53	.068
Tribunal Relação Coimbra	.48	.43	.09
Tribunal Relação Lisboa	.45	.46	.09
Tribunal Relação Évora	.44	.39	.17
p-value = 3.1e-8			

Table 2.10: Parameters of covariate Court in clusters (1-3).

Court	Cluster1	Cluster2	Cluster3
Supremo Tribunal de Justiça	-.546	-.880	1.426
Tribunal Relação Guimarães	-.475	.155	-.629
Tribunal Relação Porto	-.061	.440	-.379
Tribunal Relação Coimbra	.115	.131	-.246
Tribunal Relação Lisboa	.063	.302	-.365
Tribunal Relação Évora	-.045	-.148	.193
p-value = 3.1e-8			

Table 2.11: Parameters of statistically significant moral foundations.

Moral Foundation	Cluster1	Cluster2	Cluster3	*p*-value
Care Virtue	75.410	-104.769	29.359	3.2e-7
Care Vice	-49.525	-16.928	66.453	.002
Fairness Virtue	20.295	95.868	-116.163	.000
Fairness Vice	312.179	-258.777	-53.403	.016
Sanctity Vice	110.382	89.580	-199.962	6.7e-5

Table 2.12: Cluster probabilities of inactive covariates.

Inactive Covariates	Cluster1	Cluster2	Cluster3
Domestic Violence	.49	.36	.15
Mistreatment	.48	.37	.15
Illegal Possession of a Weapon	.29	.21	.50
Physical Aggression	.30	.54	.15
Threat	.27	.44	.28
Sexual Aggression	.42	.11	.48
Homicide	.11	.12	.77

2.5.2. CHARACTERIZATION OF THE THREE LATENT CLUSTERS

Affirmative with suspended prison time This cluster relates to gender-based violence cases that resulted in an affirmative legal outcome (accusation, conviction, coercive or safety measures), but, when such outcome entails prison time, offenders are often given a suspended sentence.

According to Portuguese Law, a suspended sentence is possible in cases featuring a conviction not exceeding five years of prison time, if the court considers that the threat of imprisonment adequately and sufficiently fulfills the purposes of punishment. Offenders in this cluster's cases have a higher probability of remaining in freedom, provided they do not re-offend and comply with the rules of conduct established by the court.

In this cluster, several offenses emerge along with domestic violence and mistreatment. We observe a prevalence of offenses related to physical aggression, sexual aggression, and illegal possession of a weapon (dataset).

The cases featured in this cluster are well distributed among the six Portuguese appeal courts included in this study, although we report a slightly higher probability of the cases being decided in the Tribunal da Relação de Guimarães.

The moral rhetoric in the cases of this cluster is rooted in foundations of care virtue, fairness vice, and sanctity vice. When estimating the membership probabilities for each cluster, taking into account different values of these foundations while holding the other foundations at their mean values, we observe that, as the frequency of the moral foundations

increases, there is also an increased probability of membership in this cluster (Figures 2.3,2.4, 2.5).

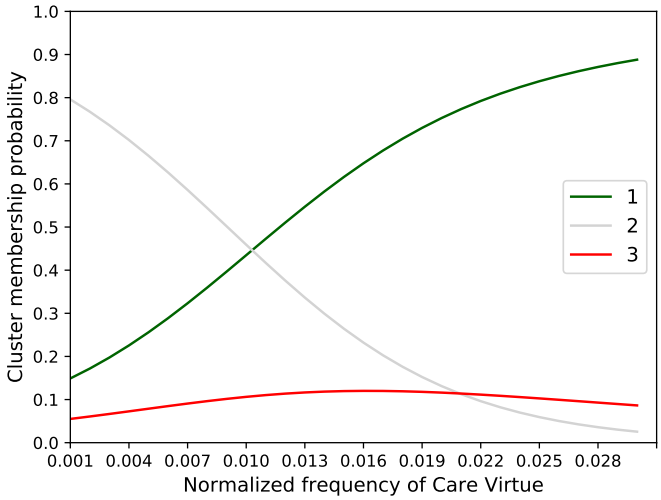


Figure 2.3: Cluster membership prediction: care virtue.

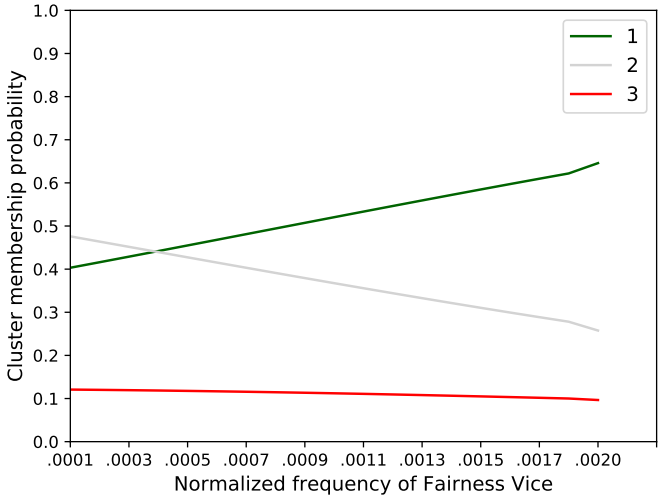


Figure 2.4: Cluster membership prediction: fairness vice.

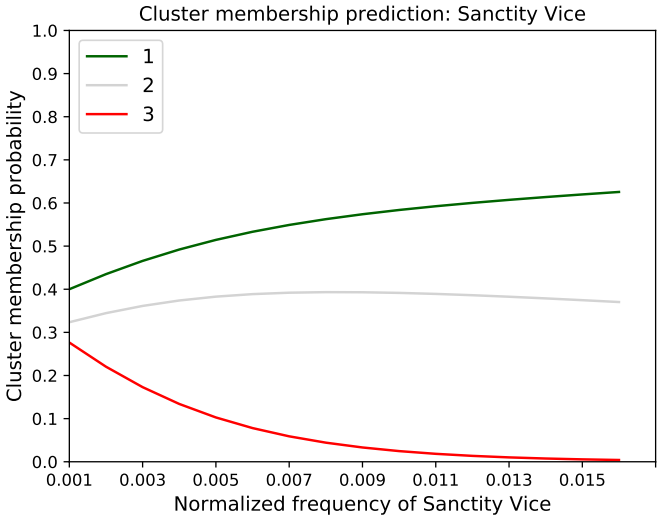


Figure 2.5: Cluster membership prediction: sanctity vice.

Mixed outcomes but no prison time This cluster relates to a mix of gender-based violence cases featuring affirmative, not affirmative, and neutral legal outcomes but no prison time.

It is not surprising that cases that are either not affirmative of gender-based violence or neutral, as they relate to procedural matters during the investigation or trial, do not entail prison time as there is no conviction.

This does not hold entirely for the cases that are affirmative of gender-based violence, which are also prevalent in this cluster. Some of those cases relate to ongoing legal investigations in which the court makes a formal accusation and applies a coercive measure to the suspect offender, but there is no conviction as the trial is still pending. Other cases relate to final investigations which resulted in a conviction, but the offender is sentenced to a court fine rather than prison time.

In this cluster, along with domestic violence and mistreatment offenses, there is also a pronounced prevalence of offenses related to physical aggression, threat, and illegal possession of a weapon (dataset). The affirmative cases resulting in a court fine which were mentioned above, often relate to offenses of physical aggression and threat.

The cases featured in this cluster are well distributed among the five intermediate appeal courts, but there is a lower probability of these cases being decided in the Supreme Court.

The moral rhetoric in these cases indicates fairness virtue considerations. When estimating the membership probabilities for the clusters as described above, we observe that, as the frequency of fairness virtue increases, there is an upward trend with respect to the probability of membership in this cluster (Figure 2.6).

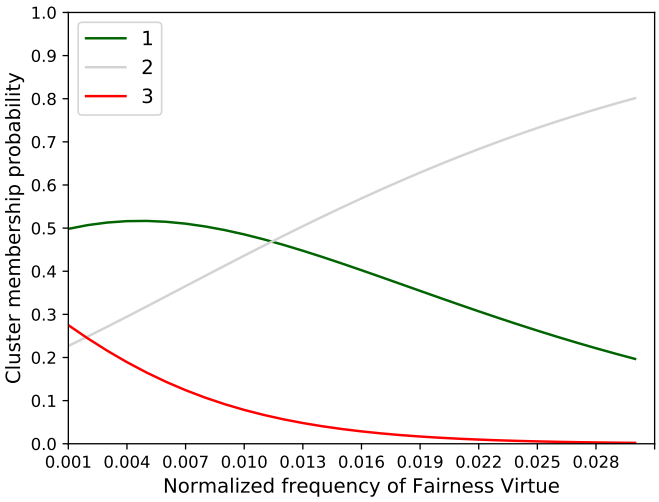


Figure 2.6: Cluster membership prediction: fairness virtue.

Affirmative with much prison time and compensation This cluster relates to gender-based violence cases with affirmative legal outcomes that resulted in long prison time and high monetary compensation.

These are complex cases featuring a multitude of psychological, sexual, and physical abuse that sometimes result in the loss of lives. Along with domestic violence and mistreatment, there is a prevalence of homicide, illegal possession of a weapon, and sexual aggression offenses in this cluster (dataset). Given the nature of these offenses, it is not surprising that these legal cases are most often decided in the Supreme Court of Justice.

The moral rhetoric in the cases featured in this cluster is mainly focused on care vice. When estimating the membership probabilities for the clusters as described above, we observe that, as the frequency of care vice increases, there is an upward trend with respect to the probability of membership in this cluster (Figure 2.7).

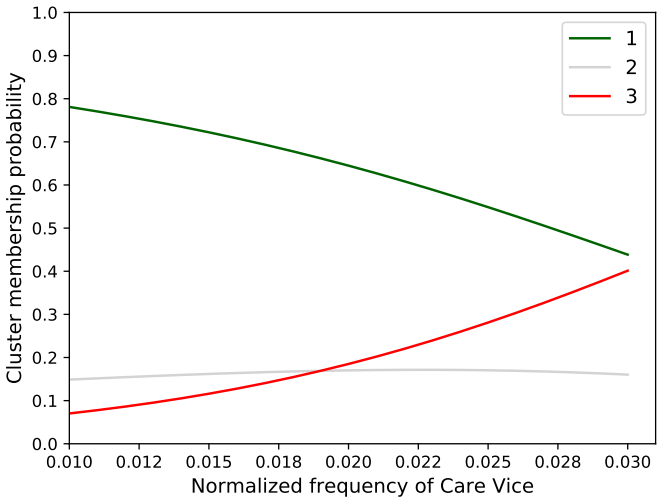


Figure 2.7: Cluster membership prediction: care vice.

2.5.3. INDIVIDUALIZING VS. GROUP FOUNDATIONS

Our results indicate that the moral rhetoric in the gender-based violence legal cases of this study is dominated by individualizing moral foundations (Table 11). Except for sanctity vice, all statistically significant moral foundations in this study are individualizing (care and vice). This means that the moral rhetoric in these cases is dominated by considerations of protecting individuals and not so much about the protection and well-being of collectives.

2.6. DISCUSSION

In this study, we aimed to investigate the relationship between moral foundations and morally relevant behavior. Rather than relying on self-scoring and social experiments, which have well-documented shortcomings, we focused on legal narratives, which tend to be morally charged and have an associated behavioral outcome. Our case study is on gender-based violence legal cases decided in Portuguese (appeal) courts.

The research is anchored in domestic violence, but we also accounted for other offenses emerging in these legal cases. Offenses related to illegal possession of a weapon and physical aggression are prevalent in every cluster. In contrast, offenses of sexual aggression are prevalent in *Affirmative with suspended prison time*, offenses of threat are prevalent in *Mixed outcomes but no prison time*, and offenses of homicide are prevalent in *Affirmative with much prison time and compensation*. These offenses are core elements of the social problem of gender-based violence in Portugal.

This study contributes to MFT research by further exploring the relationship between moral foundations and behavior. The clusters identified in the data unravel latent relations between the moral rhetoric in gender-based violence legal narratives and corresponding legal

outcomes. The moral nuances found in this study require further considerations in light of the moral foundations previously contextualized within gender-based violence.

Our results show that the care foundation is associated with the clusters in which the gender-based violence offender is often convicted to prison time. The moral rhetoric in these cases is imbued with intuitions of protection of vulnerable groups, such as women and children. These intuitions are therefore associated with the decision of the court to sentence the offender to prison time, although the causal mechanics of such association remains unclear.

The different dimensions of the care foundation are associated with the severity of prison sentences. In the *Affirmative with suspended prison time* cluster, in which the offenders are often given a suspended prison time sentence, the moral rhetoric relies on the virtue side of the care foundation. This may indicate that there are also care considerations about rehabilitating the offender. In the *Affirmative with much prison time and compensation* cluster, in which the offenders are often sentenced to long prison time, the moral rhetoric relies on the vice side of the care foundation. This may indicate that the care considerations are focused on the victims and there is an emphasis on the vice behavior of the offender.

Fairness considerations are also found in the narratives of gender-based violence cases. The fairness foundation is prominent in clusters where offenders spend little or even no time in prison. In *Affirmative with suspended prison time* the court often makes a decision to suspend the prison sentence, thus the fairness foundation may operate as a moderating factor of punitive responses to gender-based violence. However, when looking at the fairness dimension prevalent in this cluster (fairness vice), it seems that there are also strong considerations about the unfair, exploitative, and deceiving behavior of the offender. In *Mixed outcomes but no prison time* there are often procedural decisions related to the secrecy of the investigation, revision of legal decision, or retrial. Thus, the fairness foundation may operate as an element of due processing. The fairness dimension that is prevalent in this cluster (fairness virtue) does indeed relate to equality, civil rights, proportionality, and due processing of Justice.

Our results also show sanctity considerations in the narratives of gender-based violence cases. We had previously reasoned that this moral foundation may be associated with cases featuring sexual offenses. In this study, we found that it is associated with *Affirmative with suspended prison time*, which includes several cases of sexual aggression. However, we did not find an association with *Affirmative with much prison time and compensation*, which also features sexual aggression. This may be explained by the fact that cases in the latter cluster are mostly decided in the Supreme Court of Justice, which tends to focus on the formal and procedural matters of legal cases. Therefore the narratives may be less conducive to sanctity rhetoric.

We report a dominance of individualizing moral foundations in the narratives of gender-based violence legal cases. Individualizing foundations (care and fairness) elicit moral concern for individuals, whereas binding moral foundations (loyalty, authority, sanctity) elicit concern for the needs and well-being of collectives. Except for sanctity/vice, all moral foundations found statistically significant in this study are individualizing. The endorsement of individualizing and binding foundations has been associated with variations in punitive responses.

Findings reported in the literature suggest that the endorsement of individual-oriented foundations is associated with lower punitiveness, whereas binding-oriented foundations are associated with higher levels of punitiveness [198, 203, 225]. This does not seem to be the case in our study, as the prevalent foundation in the cluster *Affirmative with much prison time and compensation* is care/vice. However, if we assume that the cases decided by the Supreme Court focus mainly on the formal and procedural matters of legal cases and we compare only the clusters featuring cases most often decided on the intermediate appeal courts (*Affirmative with suspended prison time* and *Mixed outcomes but no prison time*), we report that the only binding foundation that is statistically significant in this study (sanctity/vice) is associated with the cluster featuring higher levels of punitiveness (*Affirmative with suspended prison time*). Further research is required to include decisions of lower-level courts and understand the rhetoric nuances between the different levels of court.

In addition to the contributions to MFT research, this study also contributes to gender-based violence research. We further reflect on our case study about gender-based violence in Portugal. It is important to emphasize that each data point in this study is not just a number but a tragic situation involving the lives of human beings. Important overarching considerations about the clusters identified in the data open avenues for future research and policy endeavors.

The offenses consistently emerging in the clusters along with domestic violence signal that physical aggression and illegal weapons are core elements of gender-based violence in Portugal. The distinction between domestic violence and physical aggression, which rests on the idea that physical aggression is related to the physical well-being of a person, whereas domestic violence is a broader concept that entails physical, psychological, and mental well-being, is often blurred and difficult to prove in court. Further research is needed to understand if these cases are being correctly framed within the legal system, the relation between physical aggression and domestic violence, i.e., whether physical aggression situations evolve toward domestic violence, and which strategies are more efficient to tackle violence at the earliest sign.

Given the prevalence of offenses related to illegal weapons, further research is needed to understand the motivations and chain of events that lead gender-based violence offenders to acquire illegal weapons.

This study also provided the opportunity to investigate gender bias in the narratives and decisions of domestic violence legal cases. As mentioned earlier, this is a much-debated topic in Portugal in the aftermath of a controversial legal decision that cited the Bible and an 1884 Penal Code to frame a case of domestic violence. We identified narratives that reflect outdated and biased sociological views on gender-based violence [11], but in this study, we have not found the gender of judges to be statistically associated with legal outcomes. The moral rhetoric rooted in individualizing foundations also indicates that gender-based violence legal narratives focus on protecting vulnerable groups and progressive rehabilitation-oriented policies that aim to integrate offenders in the community successfully.

Our study has important limitations that should be accounted for. The first limitation

[11] ECLI: PT:STJ:2003:03P2394.EE and PT:TRL:2013:1354.10.6TDLSB.L1.5.9F.

concerns the data used in the study. As mentioned in the methodology section, we used the Jurisprudence Research Portal to retrieve the gender-based violence legal cases. This is a very comprehensive platform, yet it does not feature all cases decided in the Portuguese courts. It does not include the cases decided in the lower courts, which are also not featured in our study. This is an unfortunate limitation as the narratives in higher courts are often limited to matters of appeal. Moreover, the Jurisprudence Research portal only features a sample of the legal cases decided in the appeal courts, and there is no information available about the sampling criteria. Finally, we also identified 90 cases in the Jurisprudence Research portal where only the name of the judge presiding the court was featured in the legal decision, and there was no information about the remaining judges involved in the (collective) decision. This limitation did not allow us to investigate the variations in conviction rates when judges are deciding individually, which have been reported to be higher for female judges in a recent study [209].

A second limitation is related to moral rhetoric. In this study, we assume that the judges compose the legal narratives in a way that supports their views on the case. We acknowledge, however, that a legal decision is a complex narrative featuring legal jargon, statements from the parties involved, and the decision of the court. It is therefore challenging to untangle these different pieces and clearly identify the moral rhetoric of the judge leading to the legal decision. Moreover, the causal mechanisms remain unclear. This is a well-known limitation in MFT research. It is unclear whether the moral foundations lead to the legal decision or the legal decision leads to the moral foundations.

A third limitation concerns our conceptualization of a moral decision as morally relevant behavior. In this study, we investigated the relationship between moral foundations and morally relevant behavior in gender-based violence legal cases. We assumed that a legal decision is a morally relevant behavior. However, we acknowledge that institutional and legal limitations severely constrain such decisions.

Future research may address the limitations outlined above and extend this study by including a sample of legal cases decided in Portuguese lower courts, exploring ways to untangle the moral rhetoric in legal cases, for instance, through NLP, and using different dictionaries to map the moral rhetoric. An interesting research avenue would be to survey judges who made legal decisions included in this study on the MFQ and compare those moral scores with the moral rhetoric in legal cases. By comparing those scores, we would further understand the differences between the moral profile of judges and how they compose morally charged narratives in legal cases.

2.7. CONCLUSION

The MFT is a theoretical model of intuitive ethics that builds on the assumption that a small number of moral foundations related to different adaptive challenges emerge as intuitive reactions to social stimuli. Recent MFT research has explored the relationship between moral foundations and human behavior. However, these studies rely on self-scoring and social experiments, which have well-documented limitations. In this study, we expanded this line of research by investigating the relationship between moral foundations in gender-based violence legal narratives and corresponding legal outcomes.

We used a Latent Class Cluster Analysis model to identify the complex patterns in the data

and reduce such patterns to a distinct number of clusters. Three main clusters unravel latent associations between the moral foundations mapped in the narratives and legal outcomes: (i) *Affirmative with suspended prison time*; (ii) *Mixed outcomes but no prison time*; and (iii) *Affirmative with much prison time and compensation*.

The findings of this study show a dominance of individualizing moral foundations in our sample of gender-based violence legal narratives. The care foundation is associated with clusters where the offender is often convicted to prison time. The fairness foundation is associated with clusters in which the offenders spend little or no time in prison due to procedural decisions related, for instance, to suspended sentences or re-trial. When adjusting the level of courts at the intermediate appeal level, the sanctity foundation is associated with the cluster featuring sexual offenses.

This study contributes to MFT research by exploring the relationship between moral foundations and behavior. We established that there is an association between moral foundations in gender-based violence narratives and corresponding legal outcomes, which we conceptualize as morally relevant behavior. Further research is needed to understand better the causal mechanisms underlying this association.

II

AI as a System Operating in a Morally Charged Context

3

Ethical Issues in Focus by the Autonomous Vehicles Industry

A. Martinho, N. Herber, M. Kroesen, and C. Chorus, *Ethical issues in focus by the autonomous vehicles industry*, Transport reviews **41**, 556 (2021).

Abstract

The onset of autonomous driving has provided fertile ground for discussions about ethics in recent years. These discussions are heavily documented in the scientific literature and have mainly revolved around extreme traffic situations depicted as moral dilemmas, i.e. situations in which the autonomous vehicle (AV) is required to make a difficult moral choice. Quite surprisingly, little is known about the ethical issues in focus by the AV industry. General claims have been made about the struggles of companies regarding the ethical issues of AVs but these lack proper substantiation. As private companies are highly influential on the development and acceptance of AV technologies, a meaningful debate about the ethics of AVs should take into account the ethical issues prioritised by industry. In order to assess the awareness and engagement of industry on the ethics of AVs, we inspected the narratives in the official business and technical reports of companies with an AV testing permit in California. The findings of our literature and industry review suggest that: (i) given the plethora of ethical issues addressed in the reports, autonomous driving companies seem to be aware of and engaged in the ethics of autonomous driving technology; (ii) scientific literature and industry reports prioritise safety and cybersecurity; (iii) scientific and industry communities agree that AVs will not eliminate the risk of accidents; (iv) scientific literature on AV technology ethics is dominated by discussions about the trolley problem; (v) moral dilemmas resembling trolley cases are not addressed in industry reports but there are nuanced allusions that unravel underlying concerns about these extreme traffic situations; (vi) autonomous driving companies have different approaches with respect to the authority of remote operators; and (vii) companies seem invested in a lowest liability risk design strategy relying on rules and regulations, expedite investigations, and crash/collision avoidance algorithms.

3.1. INTRODUCTION

The onset of autonomous driving has provided fertile ground for discussions about ethics in recent years. In addition to the ongoing debates regarding ethical issues particular to automated driving systems-equipped vehicles, the disruptive yet mundane nature of this technology dictated its archetypal role in broader conversations about ethics and Artificial Intelligence [161, 226, 227]. The vehicle at the center of these discussions is a machine designed to provide conveyance on public streets, roads, and highways with conditional, high, or full driving automation [228]. Such a machine is commonly known as Autonomous Vehicle (AV)[1] and is expected to yield a multitude of social benefits to traffic safety, mobility, and accessibility [72, 76, 231].

The ethics discussions associated with AVs which are documented in the scientific literature have mainly revolved around extreme traffic situations depicted as moral dilemmas, i.e. situations in which the AV is required to make a difficult moral choice between actions in traffic which will result in different combinations of lives saved and sacrificed [133, 134, 232, 233]. Scholars have debated exhaustively the relevance of the AV moral dilemma [140, 142, 234], the merits of using different ethical frameworks, such as Deontology, Utilitarianism, or Rawlsianism, as control algorithms for AVs [144, 145, 147, 235, 236], and the moral preferences and societal expectations about the ethics to be encoded in AVs [133, 134]. Other ethical issues have been addressed in the scientific literature, such as ethical design, accountability, human meaningful control, sustainability, and privacy [237–242].

Quite surprisingly, little is known about the ethical issues in focus by the AV industry. General claims have been made about the struggles of companies over such issues but these lack proper substantiation [134, 243, 244]. As private companies are highly influential on the development and acceptance of AV technologies [245], their stance on ethics should be taken into account for the purposes of a meaningful debate about the ethics of AVs.

In order to assess the ethics awareness and engagement of industry, we inspected the narratives in official business and technical reports of companies operating in the AV field. In this research we focused on the companies with an AV testing permit in California, where there was an early adoption of comprehensive regulations governing the testing of AVs [95, 246]. We believe that the analysis of reports from a wide range of technology and manufacturing companies in the forefront of AV technology allows us to draw important insights about ethics within the AV industry.

We first provide an overview of the ethics narratives both in the scientific literature and industry reports. At this point it should be noted that the main aim of this paper is not to present an exhaustive review of the scholarly literature concerning ethical issues surrounding the development and deployment of AVs. Rather, our aim is to explore how the discussion of ethical issues in industry reports and its counterpart in the academic literature relate and compare to one another.

[1] Different nomenclatures are used for highly automated vehicles such as *autonomous vehicles*, *automated vehicles*, *self-driving cars*, or *driverless cars* [229]. Here we adopt *autonomous vehicles* when referring to automated driving systems-equipped vehicles (levels 3, 4, or 5 driving automation systems according to the Society of Automotive Engineers International Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles [228]) for reasons of consistency with the nomenclature favored by the industry. In this context *autonomy* is associated with the ability of a vehicle to determine its operational environment, thus modulating its behavior according to relevant norms, needs or constraints [230].

For reasons of brevity, we focus on the matters of *safety and cybersecurity*, *accountability*, and *human oversight, control, auditing of AVs* as presented in the scientific literature thus raising critical yet practical questions for which we will look for answers in the industry narratives. These three issues, which we selected as our focus points, have generated a particularly rich debate in both streams of literature, and are often discussed in relation to one another. We expect that, by providing empirical insights from industry, we can make a contribution for a richer, less speculative, and more meaningful debate on the ethics of AVs.

3.2. METHODOLOGY

The ambiguous nature of ethics makes systematization challenging. Here, we attempt to alleviate ambiguity by building our research around a list of ethical issues compiled from major guidelines of AI ethics [152]. We use this list of AI ethical issues to guide us in identifying the ethics within the scientific and industry narratives.

Interestingly, Hagendorff is reluctant about the effectiveness of AI ethics guidelines. He argues that these sorts of guidelines, traditionally based on a deontological approach to ethics which relies on fixing a set of principles and maxims, should be augmented with a virtue ethics oriented approach aiming at addressing values, attitudes, and behavioral dispositions that would ultimately help professionals refraining from unethical actions [152]. We acknowledge the limitations of these deontology-based ethics guidelines in promoting a robust ethics culture within organizations. And moreover it is noted that, because we are using a list of ethical issues based on deontological guidelines as a guidance tool in this research, our results will necessarily reflect such top-down deontological approach, thus leaving out other potential relevant ethical approaches and principles related, for instance, to informed consent and risk acceptance [247].

While acknowledging these limitations, we believe this list is adequate for our research given that it includes a comprehensive and state-of-the-art compilation of ethical issues in the field of AI ethics. The original list featuring 22 ethical issues in published guidelines about AI [152] was adjusted for this research. We removed one ethical issue ("field-specific deliberations") as well as all "AI" references as we focus on the AV as a particular AI-powered technology. The final list of 21 ethical issues can be found in Figure 3.1.

We first reviewed the scientific literature, with the aim of outlining the AV ethics debates, by identifying the ethical issues prioritized by the scientific community and the main empirical findings. For this purpose we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework [248]. Using the combination of keywords “Autonomous Vehicles” AND “Ethics”; “Autonomous Vehicles” AND “Moral”; “Self driving” AND “Ethics”; “Self driving” AND “Moral”; “Driverless” AND “Ethics”; and “Driverless” AND “Moral” in Scopus and Google Scholar (in the latter only records within the 2015-2020 timeframe in the first five pages of the database were considered), 715 records were identified. Five additional records were identified through other sources, such as citation chaining. Upon initial screening of the 720 records, 324 duplicates were removed, which meant that 396 records were assessed for eligibility. Only published scientific documents with available full text written in English language and addressing the moral or ethics dimensions of automated driving systems were considered eligible for this study. Therefore, based on this eligibility criteria, 158 records were excluded (23 records were not published scientific

documents; 47 records did not have full text available; 3 records were not written in English language; and 85 records were not about the moral or ethics dimensions of automated driving systems) thus leaving a total of 238 records for further review and analysis. It is recalled that, as far as this literature review is concerned, our aim is quite modest as we intend solely to provide an outline of the AV ethics debates in the scientific literature. We started by reading and analyzing the articles in order to identify the ethical issues prioritized by the scientific community. Thereafter, we divided the articles into theoretical (195) or empirical (43), depending on the type of research employed. And finally we reflected on the theoretical propositions and main empirical findings related to the ethical issues which are the focus points of this research.

Following the review of the scientific literature, we proceeded to the document review of AV business and technical reports, relying on the list of ethical issues mentioned above, to identify the relevant issues within the industry narratives. For the selection of companies, we used the record of companies with an AV testing permit both with and without a driver in California made available by the Department of Motor Vehicles as of June 2020. The technical and business reports from the past five years were requested from the companies and also screened through standard online searches on their websites. For reasons of reproducibility of this research we only considered reports that could be downloaded and saved as portable document format (pdf) files thus excluding articles, blog entries, or other materials made available by the companies online but which carried the risk of not being accessible in the future. In total we used 86 documents from 29 companies.

These reports are curated documents that serve the purpose of communicating corporate information to investors, consumers, and regulatory agencies. For that reason, such reports may depict augmented or abbreviated accounts of the range of actions taken by AV companies with respect to ethics. These documents are, nevertheless, important pieces of information to learn the industry's formal stance on the complex ethical issues associated with AVs.

Initially, we proceeded with the reading of the documents and selection of statements that signaled ethical considerations in the context of AVs. Although contextual analysis is crucial for this investigation, we acknowledge the limitations of the manual approach. Therefore, on a second occasion, we relied on linguistic-based text data analytics in order to assess the validity of our initial results. We started by creating lexicons, i.e. groups of search keywords organized to investigate a concept [249], associated with each ethical issue. Subsequently, we applied a text mining algorithm using the previously created lexicons as regular expressions in order to locate the keywords associated with each ethical issue in the 86 documents. The output generated by this algorithm is a report stating the number of occurrences of the keywords in the lexicons associated with each ethical issue in each one of the documents issued by the AV companies. And lastly we compared the results of the text mining algorithm and the manual approach and made the necessary adjustments with reference to the contexts of the narratives. For the quantitative analysis of the results we did a standard descriptive statistics analysis of the ethical issues found in the AV industry reports [2].

[2] Information regarding the companies with a testing permit in California, the reports used in this study, and the lexicons is available in the dataset stored in the 4TU. Center for Research Data in doi:10.4121/13348535.

3.3. OVERVIEW OF THE ETHICS OF AV TECHNOLOGY IN SCIENTIFIC LITERATURE

The amount of attention that ethics of AV technology has received in recent years is quite new to the field of Transportation. Traditionally, the ethics debates in this field have revolved around less sensational issues, such as cost-benefit analysis of transport projects or fairness in pricing [250]. The advent of autonomous driving is a remarkable scientific and engineering achievement that has given rise to novel and controversial ethical issues.

Our review showed quite clearly that the scientific literature on AV ethics is dominated by considerations about *safety and cybersecurity* concerning the programming of extreme traffic situations. This controversial issue is commonly known as the *trolley problem* in reference to a thought experiment popularized by Philippa Foot in 1967 in which an agent needs to make a difficult choice of allowing a runway trolley to proceed its course and kill five track workers or divert the trolley from its course killing only one worker [251, 252]. There are many variations and extensions to this thought experiment but its core can be defined as a moral choice between actions in traffic which will result in different combinations of lives saved and sacrificed. Because extreme traffic situations need to be programmed in advance, AV technology seemed to bring this textbook thought experiment to life thus capturing the attention of scholars and the media.

We found references to the trolley problem in more than half of the 238 reviewed articles. Most of these articles are theoretical pieces of research, often written as argumentative or normative essays, about different perspectives and dimensions of the AV moral dilemma. While these debates are certainly very rich, we found this stream of the literature to be quite fragmented. For instance, there is still little consensus about the relevance of the trolley problem in the context of AVs [140, 142, 148, 163, 234].

The empirical findings reported in the literature are also quite controversial, as they reveal potential challenges in adapting societal expectations to moral decision making driving algorithms [253]. The AV social dilemma, i.e. a conflict between individual and collective interest in the context of autonomous driving technology, illustrates such challenge. It has been reported that people approve and would like others to buy utilitarian AVs which sacrifice their passengers for the greater good, yet prefer to ride in AVs that protect their passengers at all costs thus disapproving utilitarian regulation of AVs [133, 235, 254].

A substantial amount of research, namely the Moral Machine Experiment (MME) [134], has focused on collecting and analyzing moral preferences and societal expectations about the ethics to be encoded in AVs. However, the methodological soundness and value of such investigations for the purpose of defining moral algorithms for AVs have been questioned [150, 255]. The main contribution of the MME, regarding moral preferences in AV moral dilemmas, conflicts with current ethical guidelines, such as rule 9 of the German Ethics Code for Automated and Connected Driving, which prohibits distinctions based on personal features in the case of unavoidable accident situations [253, 255, 256]. The proponents of the MME acknowledge that AV policy should not necessarily follow public expectations and preferences but they believe that such preferences should not be completely dismissed. They argue that, given the strong preference for sparing children, it would be challenging to explain the rationale for not assigning a special status to children [134]. Recently, it has been hypothesized that AV fatalities carry more weight because those are rare events and

not so much due to intrinsic differences in public perception between AV and conventional vehicles fatalities [257].

Several scholars have reported an overstatement of the *AV trolley problem* and called for the ethics community to focus on other ethical issues associated with AVs [140, 258, 259].

Other issues debated in the literature include ethical design, accountability, human meaningful control, sustainability, and privacy [237–242]. Particularly, matters related to accountability and human meaningful control have received considerable attention in the literature recently and, along with safety and cybersecurity, will be be further explored later in this research.

3.4. OVERVIEW OF THE ETHICS OF AV TECHNOLOGY IN AV INDUSTRY REPORTS

In this research we focus on the AV industry in California, a State that has been an early and strong proponent of this technology and hosts many R&D programs [260]. As of June 2020, the California DMV had listed 66 permit holders for testing with a driver, 2 permit holders for driverless testing, and 0 permit holders for AV deployment.

For our analysis we used a total of 86 documents issued by 29 companies[3] in the forefront of AV technology. Therefore, we consider their official reports as important pieces of information about the industry's formal stance on the complex ethical issues associated with AVs.

Overall, the AV reports reviewed in this research show a overwhelmingly positive tone about AV technology, which needs to be interpreted in light of such reports being written for a particular audience of investors, consumers, and regulatory agencies. Although lacking the rigor and depth of the narratives in the scientific literature, a plethora of ethical issues are referenced in the AV industry reports.

A quantitative synopsis of the ethics prioritized by companies with an AV testing permit in California, based on the number of companies that addressed each one of the 21 ethical issues in their reports, can be found in Figure 3.1. It shows that *safety and cybersecurity* is the ethical issue that was addressed by more companies, followed by *common good, sustainability, well-being*; *human oversight, control, auditing*; and *science-policy link*. In contrast, issues such as *hidden costs*, *fairness, non-discrimination, justice*, or *responsible research funding* were not addressed in these reports.

[3] Waymo LLC; Tesla Motors; Nissan; BMW; Ford; Valeo North America Inc.; AutoX Technologies Inc.; Nuro Inc.; Apple Inc.; TuSimple; Aurora Innovation; Toyota Research Institute; Intel Corp; TORC Robotics Inc.; EasyMile; Ridecell; Mercedes Benz; Bosch; GM Cruise LLC; Honda; Zoox Inc.; NVIDIA Corporation; Navya Inc.; Udelv; Pony.AI; Continental Automotive Systems; Mando America Corporation; Uber Advanced Technologies Group; and AImotive Inc.

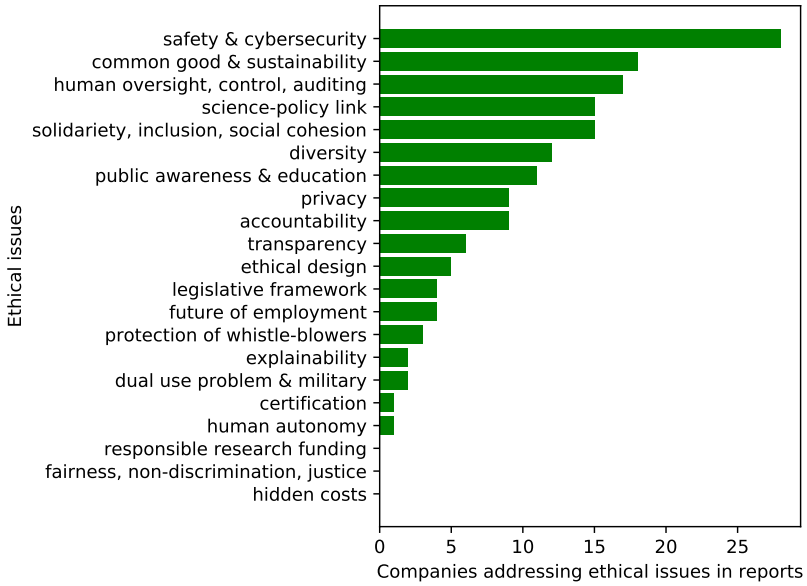


Figure 3.1: Ethical issues in AV industry.

The relevance of particular ethical issues within industry narratives requires a balance between frequency and comprehensiveness. We recall that three ethical issues had previously been selected from the scientific literature to be further explored in this research (*safety and cybersecurity*, *accountability*, and *human oversight, control, auditing*). Safety and human oversight issues are frequently addressed by AV companies and, whereas accountability is addressed by a lower number of companies, it is comprehensively explored in the reports we analyzed.

It should be noted that, even when we found quite comprehensive accounts on particular ethical issues, the narratives in the AV reports are consistently pragmatic and oriented towards technical solutions. For instance, in its reports, Mercedes-Benz acknowledges the importance of data privacy while emphasizing practical data-protection-friendly solutions that provide privacy by design in compliance with privacy laws (Daimler Sustainability Report 2018 and Reinventing Safety: A Joint Approach to Automated Driving Systems).

3.5. SAFETY AND CYBERSECURITY

Safety and security are both related to the integrity of systems but, whereas *safety* concerns the adequate functioning of a system, *security* is about the ability of a system to resist intentionally malicious actions. There are unsettled considerations about the acceptable safety and cybersecurity levels of AVs, both in mundane and extreme situations, in order to secure the well-being of users and other traffic agents [261–263].

SAFETY AND CYBERSECURITY OF AVS IN THE SCIENTIFIC LITERATURE

MUNDANE AND EXTREME TRAFFIC SITUATIONS

Mundane traffic situations are the day-to-day interactions of traffic agents (pedestrians, cyclists, animals) that require some flexibility, such as crossroads, highway entrances, or crosswalks with limited visibility. These interactions are challenging for AVs not only because these systems lack human intuition and flexibility but also because of the large scale fleet programming that is needed [140, 264]. Coordination over different technical approaches to these traffic situations is crucial to ensure safety but it is unclear how such technological coordination can be facilitated in the competitive AV market [140].

Extreme traffic situations are unexpected occurrences in the traffic environment which entail danger for vehicle occupants and other traffic agents (e.g. the unexpected appearance of an animal on a highway). Some of these situations are depicted in the scientific literature as moral dilemmas. As mentioned above, these difficult moral situations are highly explored in the scientific literature with reference to the *trolley problem* thought experiment [251, 252].

The underlying argument in the debates that take place in the scientific literature about the *AV trolley problem* can be outlined as follows: (i) *AVs ought to save lives.* (ii) *However, upon deployment of AVs, extreme traffic situations will not be completely avoided.* (iii) *Some of these extreme traffic situations will require AVs to make difficult moral decisions.* (iv) *Difficult moral decisions in traffic resemble the trolley problem.* (v) *The best option to assist AVs in managing the AV trolley problem is x.* (vi) *Option x is programmable.* (vii) *Therefore AVs should be programmed with x.* The disputes in the literature about the AV trolley problem are mainly related to premises (iv) regarding the relevance of the trolley problem in the AV context and (v) regarding the merits of different approaches to assist AVs in moral decision making in extreme traffic situations.

RELEVANCE OF THE TROLLEY PROBLEM

The relevance of the trolley problem in the AV context concerns its value as a model to investigate a relevant AV ethical challenge. It has been acknowledged that using trolley cases as inputs for crash optimization algorithms invites a myriad of criticism [234, 265]. Scholars have argued that trolley cases are of limited usefulness for the ethics of AVs because such cases would not only be highly improbable occurrences, but also their assumptions are unrealistic (outcomes of the different moral decisions available to the agent are known rather than probabilistic), inconsistent (agent has control over a vehicle yet a collision is imminent and unavoidable), and limited with respect to design (trolley cases assume a top-down approach in which an agent makes a decision explicitly, thus failing to encompass different design approaches to decision making) [140, 266].

It is accepted that trolley cases are dramatic, stylized, black-and-white situations that have little resemblance to real life extreme traffic situations. However, it is also widely acknowledged that AVs will not eliminate crashes [72, 97, 163, 267]. Therefore, weak trolley cases seem to be plausible. An example of a weak trolley case from the literature is an AV which is traveling across a two-lane bridge when a bus in the other lane swerves into its lane and the AV needs to decide either to brake, which would result in a collision with the bus, or to swerve into the other lane, thus hitting the side of the bridge [268]. Such extreme traffic situations, entailing decisions about who is put at marginally more risk of being sacrificed, may be rare occurrences when AVs are deployed but they need to be addressed

[163].

Recently, it has been proposed that the relevance of the trolley cases in the AV context is associated with the prospect of development of novel ethical principles. These principles, formulated upon analyses of the moral intuitions that emerge in stylized cases, would ultimately guide the AV design process [234, 269].

APPROACHES TO ASSIST AVS IN EXTREME TRAFFIC SITUATIONS

Another debate in the scientific literature concerns the conflicting approaches that have been advanced by scholars to address extreme traffic situations. We recall premise (v) above, *The best option to assist AVs in managing such extreme traffic situations is x*, to clarify that in the AV ethics literature *x* tends to be proposed within the realm of Machine Ethics.

Scholars have debated the merits of using ethical frameworks such as Rawlsianism, Deontology, or Utilitarianism as the control algorithms of AVs [144–147, 235, 270] or aggregating societal moral preferences [134, 255, 271, 272] to encode ethics in AVs, thus assisting them in navigating extreme traffic situations that would require moral choices.

It was shown earlier that safety considerations are central both in scientific literature and industry reports. However, scholars have mostly debated the ethics in extreme traffic situations with reference to trolley cases. We further investigate the industry's approach to extreme traffic situations and raise two relevant questions: (i) Are extreme traffic situations resembling trolley cases addressed by industry? and (ii) What are the solutions proposed by industry to address extreme traffic situations?

SAFETY AND CYBERSECURITY IN THE AV INDUSTRY REPORTS

SAFETY AND TRUST

Considering that the commercial success of AV technology depends greatly on the trust of consumers, it is hardly surprising that the industry narratives focus mainly on safety issues. In *A Matter of Trust Ford's Approach to Developing Self-Driving Vehicles*, it is stated that *for autonomous vehicles to be accepted by the public it needs to be established that they can be trusted* [273] and in Intel's white paper *A Matter of Trust: How Smart Design Can Accelerate Automated Vehicle Adoption*, trust is also emphasized when it is stated that *before driverless AVs can be widely accepted, people must be willing to trust them with their lives and the lives of those they care about hence AVs must behave, react, and communicate in ways that make it easy for people to trust them—not only the passengers inside, but also pedestrians and the other drivers who encounter them on the road* [274].

Trust and business-related considerations may not be the only reasons for the prevalence of safety considerations in the AV reports. Autonomous driving is a complex and disruptive technology which is expected to have a major societal impact. Unlike other social and ethical issues, such as fairness or human autonomy in a AI-dominated society, safety challenges are prone to be solved by technical or engineering approaches [152]. Therefore, AV companies tend to prioritize these issues, for which technical solutions are presented.

In the reports reviewed in this research, we found extensive safety considerations both for mundane and extreme traffic situations. In order for AVs to successfully deal with mundane traffic situations, companies propose advanced sensing and AI-powered solutions. Mercedes-Benz and Bosch designed a Object and Event Detection and Response (OEDR) system for AVs which is based on sensors, actuators, and computing resources

that is expected to assist the AV in handling these traffic situations - *Reinventing Safety: A Joint Approach to Automated Driving Systems* [275], whereas Valeo proposes an AI-based approach, building on the thought that in order to negotiate complex traffic conditions where there are many unknowns, AVs need to learn the data - *Meet the Future 2016 Activity and Sustainable Development Report* [276].

EXTREME TRAFFIC SITUATIONS: CRASHWORTHINESS, COLLISIONS, AND MORAL DILEMMAS

Safety considerations with respect to extreme traffic situations are also explored in the industry reports. Companies focus on the crashworthiness of AV technology, which is is quite relevant for our investigation of the AV moral dilemma as, at its core, the trolley case - either in its weak or strong version - is a convoluted crash optimization problem. By inspecting the industry reports regarding crashworthiness, we expect to clarify some of the critical elements of the AV moral dilemma.

The first element concerns the risk of crashing. Indeed the AV moral dilemma could be promptly dismissed on the account that autonomous driving will eliminate crashes. While companies express their vision of a future without accidents (*Advanced Driver Assistance Systems continues to evolve in order to realize autonomous driving and zero-accident smart vehicles, the essence of the fourth industrial revolution - Mando Sustainability Report* [277]), such ambition is mitigated by the plethora of statements on the inevitability of AV crashes and collisions which leave no room to entertain the thought of a complete elimination of accidents (*driving environments can be extremely complex and difficult and no automated driving system – regardless of how capable it may be – is likely to prevent crashes entirely - Automated Driving at Toyota: Vision, Strategy and Development* [278]; *While our top priority is to avoid collisions, we recognize it is possible that we could be involved in a collision at some point - Delivering Safety: Nuro's Approach* [279]).

Accepting that the future will not be crash and collision-free leads us to further considerations about the AV moral dilemma. Indeed crashes and collisions are a necessary condition for such extreme situations. In our document review we did not find any reference to trolley cases as described in the scientific literature, i.e. situations that require the AV to make difficult moral choices [133, 280], but we identified nuanced allusions to this matter.

Companies acknowledge that AVs will face rare extreme traffic situations, often mentioned in the industry reports as *edge cases*, and emphasize simulation and validation methods used to test these scenarios (*we test and validate our self-driving vehicles in the wide variety of environmental conditions that the vehicle might face in its operational design domain — from driving scenarios the vehicle would face daily to the rare edge cases - General Motors Self-Driving Safety Report* [281]; *AI-powered autonomous vehicles must be able to respond properly to the incredibly diverse situations they could experience, such as emergency vehicles, pedestrians, animals, and a virtually infinite number of other obstacles—including scenarios that are too dangerous to test in the real world - Nvidia Self-Driving Safety Report* [282]; *decision making is one of the most challenging tasks in the A.I. development of an autonomous vehicle...there are infinite edge cases that may be difficult or dangerous to reproduce in reality, such as illegal driving behaviors or sudden traffic accidents - The Autox Safety Factor* [283]).

We found one statement that somewhat resembles the AV moral dilemma, with one impor-

tant caveat regarding the nature of harms at stake. In the Nuro report *Delivering Safety: Nuro's Approach*, it is stated that, in the *unlikely case of a Nuro shuttle ever encountering an unavoidable collision scenario the driverless passengerless vehicle has the unique opportunity to prioritize the safety of humans, other road users, and occupied vehicles over its contents* [279]. Whereas we can not legitimately consider Nuro's account as a AV moral dilemma, we consider it as yet another indication that companies are aware of convoluted situations akin to weak versions of moral dilemmas. We speculate that Nuro's slightly more transparent stance on this matter could be explained by the fact that it focuses on passengerless self-driving delivery technology (*Our custom vehicle is engineered to make delivery of everything more accessible — from groceries to pet food, prescription drugs to dry cleaning ... with no driver or passengers to worry about, our vehicle can be built to keep what's outside even safer than what's inside... it's lighter, nimbler, and slower than a passenger car, and is equipped with state-of-the-art software and sensing capabilities that never get distracted* [279]).

At the root of extreme or edge cases are often blind spots that prevent the AV from performing an accurate evaluation of the traffic context and having enough emergency braking time. The solutions advanced by companies to address this problem rely on radars and speed limitation when the visual field of the AV is obstructed. Although blind spot detection and assistance is considered a low level automation feature, the narratives we found about this issue and its implications for pedestrians' safety are yet another substantiation of the concerns of AV companies about extreme traffic situations (*in the case of pedestrians who are occluded from the vehicle ... it should adjust the speed such that if a child would emerge from behind some object there would be no accident ... even in a worst case scenario where the pedestrian emerges from behind some sensing obstruction (e.g. a parked car) even at that maximal speed - Intel Implementing the RSS Model on NHTSA Pre-Crash Scenarios* [4] [284]; *If the view is blocked Perception will flag that area as unknown ...if an object is hard to see because of rain or fog or because it is hidden behind a truck the computer brain knows that and adjusts its decision-making and performance accordingly...this allows prudent decision-making and operation based upon both what the sensors "see" as well as what may be hidden from view* [281]).

As a result of our review of AV reports, we conclude that moral dilemmas resembling trolley cases are not addressed in these reports in the terms described in scientific and media publications, but there are nuanced allusions that unravel underlying concerns about these extreme traffic situations. Regarding the solutions proposed by industry to address extreme traffic situations, we report for now that companies rely on radars and speed limitation to address the problem of blind spots which are often at the root of these traffic situations.

3.6. HUMAN OVERSIGHT, CONTROL, AUDITING OF AVS

Human oversight, control, and auditing of autonomous systems implies the surveillance of the development and performance of the technology. It is expected that remote oversight of the performance of autonomous driving ensures trust and safety in this technology as human operators are able to take-over the vehicle. It has been reported, however, that

[4] RSS stands for Responsibility-Sensitive Safety and NHTSA stands for National Highway Traffic Safety Administration.

several technical problems take place precisely during the transfer of control over the AV [285].

HUMAN OVERSIGHT, CONTROL, AUDITING OF AVS IN THE SCIENTIFIC LITERATURE

A philosophical account of meaningful human control over automated systems has been proposed by Santoni de Sio and Van den Hoven to ground the design guidelines with respect to human oversight, control, and auditing of AV technology [242, 286]. According to this account, AVs should meet tracking and tracing conditions to allow for a meaningful form of human control [242]. An AV should therefore be able to track the relevant human moral reasons in a sufficient number of occasions, thus adjusting its behaviour in accordance to the intentions of a human designer or operator - tracking condition - and its actions should be traceable to a proper moral understanding on the part of the humans who design and deploy the system - tracing condition [242].

The tracking condition ensures that the AV complies with the intentions of a human operator. It should be noted, however, that humans are poor fallback systems. As more autonomy is added to a system and it becomes more reliable and robust, the situation awareness of human operators decreases and they are less likely to take over manual control (*automation conundrum*) [239]. Therefore, in critical situations it could be that, by meeting the tracking condition, the AV is complying with an instruction issued by a low situation awareness operator.

The tracing condition requires the presence of at least one human agent that can understand the real capabilities of the system and bear the moral consequences of the actions of the system [242]. This condition is especially relevant to tackle the responsibility gaps, i.e. situations where it is unclear who should be responsible for an outcome [287, 288], that are expected to arise in the context of AV technology as a result of the fragmentation of the technology action (*many-hands* problem [289]).

Meaningful human control has been heralded as the standard for AVs to meet the appropriate level of safety and accountability [259]. We will revisit this theory in the section below about the accountability of AVs. From this section we raise one relevant question, regarding the tracking condition, to be investigated in the industry reports: according to the autonomous driving industry, which decision prevails in traffic, the decision of the AV or the decision of the human operator?

HUMAN OVERSIGHT, CONTROL, AUDITING IN THE AV INDUSTRY REPORTS

Remote and onsite human oversight of AV operations is addressed in the industry reports analyzed in this research. We recall that few companies have a driverless testing permit in California, which means that in general companies rely heavily on onsite human oversight for the testing of AVs. *Mission Specialists are trained on the governing operational design domain, and are prepared to take manual control of the vehicle when presented with a scenario that is not included in the current operational design domain - Uber Advanced Technologies Group A Principled Approach To Safety* [290]. In addition to onsite oversight, companies also rely on remote control of AV operations. In a report issued by Zoox, it is stated that their *remote operations support center will have operators available to remotely guide vehicles at any time, day or night, when a vehicle encounters an uncertain driving*

situation such as a traffic light outage or a road obstruction - Safety Innovation at Zoox: Setting the bar for safety in autonomous mobility [291].

Building on the notion of *tracking*, introduced above in the context of Meaningful Human Control, we report different approaches with respect to the authority of remote operators. Companies such as Mercedes-Benz and Intel seem to prioritize the autonomy of the vehicle (*while automated driving vehicles take under consideration data received from an infrastructure, particularly data that can be strongly authenticated and validated, the vehicles ultimately maintain their own decision authority, not the infrastructure - Safety First of Automated Driving 2019* [292]) whereas other companies, such as AutoX, seem to prioritize the decisions made by remote operators (*operators at the remote support system can check the AI decision results and correct or overwrite them when unexpected errors occur* [283]).

The statements that we identified in the AV reports regarding human oversight relate more to the first condition of the Human Meaningful Control theory, but we also report one statement which relates to the tracing condition with respect to the understanding of the system. AImotive states that *test operators face their own unique challenges. The debug screen of a complex autonomous system is incomprehensible to the untrained eye. These engineers and developers have a deep understanding of the code at work in our prototypes allowing them, at times, to predict when the system may fail. This allows our test crews to retake control of the vehicle preemptively, in a controlled manner - Ensuring Safe Self-Driving AImotive's Development Puts Safety First* [293]. The tracing condition in the Meaningful Human Control theory has another dimension, related to responsibility, which will be addressed below.

3.7. ACCOUNTABILITY

Accountability issues associated with AV technology have received substantial attention in the scientific literature. We refer to accountability in broad terms, thus encompassing closely related concepts, such as responsibility and liability. It is clarified that accountability entails responsibility, but unlike the latter it requires explanations about actions and it cannot be shared [294]; responsibility for an action traditionally requires at least a *control condition*, i.e. an agent is responsible if it is the agent of the action, and an *epistemic condition*, i.e. awareness or knowledge of the agent regarding the action [295]; and liability is legal or financial responsibility [238]. These matters are challenging in the AV domain, mainly because of the fragmentation of the technology action, which can result in responsibility gaps.

ACCOUNTABILITY IN THE SCIENTIFIC LITERATURE

Different approaches to AV responsibility have been proposed in the literature [288, 296–298]. The theory of Human Meaningful Control, which was introduced above, encompasses a tracing condition that requires the presence of at least one human agent who can bear the moral consequences of the actions of the AV [242]. It has been asserted that, in order for the tracing condition to be met in higher order levels of automation, a transition of responsibility from the driver to designers or remote operators is required. At such levels of automation, how the AV is designed to execute its tasks is more important than how the human driver ought to execute its tasks [285].

An argument has been presented in the scientific literature particularly concerning liability, in which it is roughly stated that AVs have the potential to save lives but crushing liability may discourage manufacturers from developing and deploying AVs, and as such this technology would not meet its potential to save lives [237, 299]. As legal scholars are working on extensions to criminal and civil law [300–302], it is questioned whether liability legal frameworks should be designed in such a way that would not impede, but rather promote, the development and improvement of AVs [237]. This argument has been undermined by some scholars who claim that increased manufacturer liability will not be problematic, as AVs will be safer and will bring down the overall cost of litigation and insurance [303]. Indeed, it has been reported that thus far governments have avoided strict measures in order to promote AV developments [304].

Another issue that has been presented in the literature is the *liability dilemma of the AV manufacturer* which showcases the conflict between ethics and law when it comes to liability. When designing a crash collision algorithm, a manufacturer is assumed to face three options while balancing ethics and liability: (i) program an algorithm to swerve in a direction that would sacrifice fewer lives but would entail high liability due to compensatory and punitive damages for intentional conduct caused by targeting the sacrificed people; (ii) allow the AV to run its course which would entail a larger number of lives sacrificed but lower liability which would then be restricted to compensatory damages; and (iii) avoid a collision, which if successful, entails that that no lives are sacrificed but if unsuccessful, entails the largest number of lives sacrificed, but in either case entails the lowest liability [269]. It is therefore concluded that what is easier in a lawsuit may not be the more ethical solution [269]. By featuring a trolley case in the background and making simplistic and general assumptions about the law, the AV liability dilemma suffers from the same sort of shortcomings that have been pointed out earlier about the AV moral dilemma. Despite its limitations, the liability dilemma of the AV manufacturer sheds light on the tension between ethics and liability, which should not be ignored.

The three issues we explored above regarding the transition of responsibility in higher order levels of automation, the liability and technology development argument, and the liability dilemma can be further investigated within the industry narratives by considering the design strategies with respect to accountability. From this section we raise the question: which accountability design strategy is being adopted by the AV industry?

ACCOUNTABILITY IN THE AV INDUSTRY REPORTS: THE CASE OF THE SUPER-HUMANLY FAST RUNNER

We found several statements in the industry reports that allow us to further reflect on the approach of the industry with respect to accountability. In general, AV companies seem invested in a lowest liability risk design strategy relying on rules and regulations, expedite investigations, and crash/collision avoidance algorithms.

Companies have stated the need for clear rules to be set in advance (*It is necessary to put legal frameworks in place in order to clarify where the responsibility lies in case of the occurrence of an accident after the realization of fully automated driving - Honda Sustainability Report 2015* [305]), while also presenting technical solutions aimed at expediting investigations, such as the use of a "black-box" akin to a flight recorder designed to store accident data, or responsibility algorithms based on mathematical models (*With regard to liability*

issues a "black-box" is required that stores certain data necessary to sort out accident liability which can contribute towards allocating responsibility between human and machine when accidents happen - BMW Sustainable Value Report 2016 [306]; *What will happen when a collision occurs? There will be an investigation, which could take months ... our solution is to set clear rules for fault in advance, based on a mathematical model ... the investigation can be very short and based on facts, and responsibility can be determined conclusively. This will bolster public confidence in AVs when such incidents inevitably occur and clarify liability risks for consumers and the automotive and insurance industries - Intel A Plan to Develop Autonomous Vehicles. And Prove it.* [307]).

It was mentioned earlier that companies acknowledge that AV technology will not eliminate accidents, yet some companies are invested in developing AVs which will never cause or be responsible for accidents (*By formally defining the parameters of the dangerous situation and proper response, we can say that responsibility is assigned to the party who did not comply with the proper response. Therefore, the Responsibility-Sensitive-Safety model guarantees that when applying it to any "driving policy" (the decision-making mechanism of the AV), the self-driving car will never initiate a dangerous situation and thus, it will never cause an accident* [284]; *Over time, though, Guardian capability will grow steadily as technology improves, with a goal of creating a vehicle never responsible for a crash regardless of errors made by a human driver* [278]). In one of the Intel reports (*Intel A Plan to Develop Autonomous Vehicles. And Prove it.*) it is stated that their Responsibility-Sensitive-Safety system will always brake in time to avoid a collision with a pedestrian unless the pedestrian is running super-humanly fast [307]. By providing the super-humanly fast runner illustration, Intel is not only emphasizing that their AV will not be responsible for a collision with a pedestrian, but it is also promoting trust in their technology.

Notwithstanding the positive accounts we found in the industry reports regarding the development of minimally responsible AV technology, we found a statement in Nissan's Financial Information 2018 bracing the company for potential liability losses related to AVs: *If the autonomous driving technology is developed and its use becomes quickly widespread in the future, the responsibility of automobile manufacturers might be brought into question in connection with the decline in drivers engaged in driving...If the recalls that the Group has implemented for the benefit of customers' safety become significant in volume and amount, the Group would not only incur significant additional expenses but also experience damage to its brand image, which could adversely affect its financial position and business performance* [308].

3.8. CONCLUSION

Despite the wealth of discussions about the ethics of AVs, little is known about the awareness and engagement of the industry on this matter. In this research we have provided an overview of the narratives on the ethics of AVs as presented both in scientific literature and in industry reports issued by companies with an AV testing permit in California. Subsequently, we focused on *safety*, *accountability*, and *human oversight*, and we raised critical yet practical questions, for which we looked for answers in the industry narratives. A combination of contextual analysis and text mining techniques was employed to select statements signaling AV-related ethical considerations within the industry reports.

The overall conclusion that can be drawn from our analyses is that industry and academia look at the ethics of AV technology through rather different lenses. For example, while the scientific literature has been largely preoccupied with deep considerations of abstract moral dilemmas (trolley problem), industry reports adopt a much more pragmatic, technology-infused and perhaps overly optimistic narrative when discussing the potential of so-called edge cases where accidents cannot be avoided and loss of life and damage need to be minimized. While this discrepancy may perhaps not come as a surprise to many, it is disappointing to see that on matters that are of such great importance to the general public, science and industry seem to diverge so profoundly. While we certainly do not advise to try and establish some form of agreement between industry's views regarding the ethical issues surrounding AVs and those of academia (which would be a tall order anyway, given the wide variety of such views within industry and within academia), we do believe that it would be valuable to both sides of the aisle to inform one another of one's viewpoints.

More specifically, the findings in this research suggest that: (i) given the plethora of ethical issues addressed in the reports, autonomous driving companies seem to be aware of and engaged in the ethics of autonomous driving technology; (ii) scientific literature and industry reports prioritize safety and cybersecurity; (iii) scientific and industry communities agree that AVs will not eliminate the risk of accidents; (iv) scientific literature on AV technology ethics is dominated by discussions about the trolley problem; (v) moral dilemmas resembling trolley cases are not addressed in industry reports but there are nuanced allusions that unravel underlying concerns about these extreme traffic situations; (vi) autonomous driving companies have different approaches with respect to the authority of remote operators; and (vii) companies seem invested in a lowest liability risk design strategy relying on rules and regulations, expedite investigations, and crash/collision avoidance algorithms.

Clearly, our study has its limitations, which we would like to highlight once more at this point. Despite our efforts to alleviate ambiguity surrounding terms such as accountability, we could not successfully remove it entirely from this research. This is unavoidable as academic ethicists amongst themselves have rather diverging views on what a term like accountability means, and how it should be distinguished from related concepts such as responsibility. It should therefore not come as a surprise that this ambiguity at an abstract level may translate into different (implicit) meanings attached to the same word, in different industry reports. As a consequence, our analysis of these reports which uses a combination of 'manual reading' and text mining, risks conflating different meanings attached to the same vocabulary. One promising way to alleviate or at least diminish this problem, is to use techniques that are popular in the field of Anthropology, such as participant observation, in-depth interviews and focus groups. These techniques offer a potential window into how particular terminology is being used in the AV-industry, as such providing a base for more carefully discussing how different industry actors differ from one another in terms of their approach to, e.g., accountability in the context of AVs and how industry as a whole differs from academia in this regard.

Such techniques could also help remedy a second limitation of our study, which is that we focused on curated reports that were made publicly available by industry actors for a particular audience. Although, as we argued above, we believe that such documents hold important clues regarding the views of industry actors – e.g. in providing insight into how they like to be seen by others – there is clearly scope and need for more and other types of data collection here. For example, participant observation in which a scholar would be

allowed to be embedded in an AV-company for a longer period of time, and to do a range of in-depth interviews with employees at various levels of the organization, is likely to add significantly to our knowledge of industry's dealings with the ethical conundrums that surround the development and deployment of AVs. We trust that our study would provide a useful stepping stone for such follow up research.

4

A HEALTHY DEBATE: EXPLORING THE VIEWS OF MEDICAL DOCTORS ON THE ETHICS OF ARTIFICIAL INTELLIGENCE

A. Martinho, M. Kroesen, and C. Chorus, *A healthy debate: Exploring the views of medical doctors on the ethics of artificial intelligence*, Artificial Intelligence in Medicine **121**, 102190 (2021).

ABSTRACT

Artificial Intelligence (AI) is moving towards the health space. It is generally acknowledged that, while there is great promise in the implementation of AI technologies in healthcare, it also raises important ethical issues. In this study we surveyed medical doctors based in The Netherlands, Portugal, and the U.S. from a diverse mix of medical specializations about the ethics surrounding Health AI. Four main perspectives have emerged from the data representing different views about this matter. The first perspective (AI is a helpful tool: Let physicians do what they were trained for) highlights the efficiency associated with automation, which will allow doctors to have the time to focus on expanding their medical knowledge and skills. The second perspective (Rules & Regulations are crucial: Private companies only think about money) shows strong distrust in private tech companies and emphasizes the need for regulatory oversight. The third perspective (Ethics is enough: Private companies can be trusted) puts more trust in private tech companies and maintains that ethics is sufficient to ground these corporations. And finally the fourth perspective (Explainable AI tools: Learning is necessary and inevitable) emphasizes the importance of explainability of AI tools in order to ensure that doctors are engaged in the technological progress. Each perspective provides valuable and often contrasting insights about ethical issues that should be operationalized and accounted for in the design and development of AI Health.

4.1. INTRODUCTION

Artificial Intelligence (AI) is moving towards the health space. Given the abundance of data generated by health systems as a result of digitization efforts made over the last decade, a new data-driven approach to implement AI in healthcare has emerged. In contrast with previous and somewhat failed rule-based approaches to implement AI in healthcare [309, 310], this new approach relies heavily on algorithms that detect patterns in data from clinical practice (e.g. medical imaging and electronic health records), clinical trials, genomics studies, and insurance, pharmaceutical, and pharmacy benefits management operations [311]. There is an expectation that these state-of-the-art-data-driven AI methods and algorithms will be able to use such data to address the complex problems of health systems [311, 312].

The implementation of AI in healthcare holds great promise for expanding the medical knowledge and providing optimal yet cost-effective healthcare solutions [313, 314]. In the clinical domain, expected results include identification of individuals at high risk for a disease, improved diagnosis and matching of effective personalized treatment, and out-of-hospital monitoring of therapy response [312, 315]. Despite the projected benefits associated with Health AI, it also raises important ethical issues [316, 317].

It is well known that AI has the potential to threaten values such as Autonomy, Privacy, and Safety [152], which are core values in Medicine [318, 319]. Therefore, in order for AI to promote quality of care and minimize potentially disruptive effects [320], its deployment must take ethics into account. An important step towards ethical deployment of disruptive AI technologies is to learn the views of practitioners about such technologies. This information allows a better operationalization price2019privacyof the ethical issues associated with AI in a particular domain, which eventually is expected to lead to more meaningful debates and robust policies.

The current academic literature provides interesting and valuable information on the perspectives of practitioners about the impact of AI technologies in the medical profession [178–181]. Most of these studies are particularly suited to medical fields with a strong image processing component, which is adequate for automated analysis, such as radiology [321–327], pathology [328], and dermatology [329, 330]. However, there is little knowledge on the views of medical doctors about the ethical issues associated with the implementation of AI in healthcare.

The aim of this study is to gain insight into the reasoning patterns and moral opinions about Health AI from those involved in the medical practice. By surveying medical doctors in The Netherlands, Portugal, and U.S. on the ethical issues associated with the implementation of AI in healthcare, we expect to enrich existing literature on the impact of AI technologies in medicine and provide valuable knowledge for the operationalization of Health AI Ethics.

We first provide a brief commentary about the ethics of AI in healthcare. Subsequently we explain the methods used in this research by outlining the basic steps of q-methodology and explaining how we established these steps in this study. Later we present the results of the study by describing the four different perspectives that have emerged from the data. These results are further analyzed and discussed. Finally we draw conclusions and present directions for further research.

4.2. THE ETHICS OF HEALTH AI

The empirical work about AI in healthcare that has been reported in the literature focuses mainly on issues directly related to the medical practice and career, such as Future of Employment, Education about AI, and Accountability.

It has been reported that medical students and practitioners understand the increasing importance of AI in healthcare and have positive attitudes towards the clinical use of AI [181, 323, 329], but mainly as a supportive system for diagnosis [321, 322, 327–330]. Despite the positive attitudes towards AI, it has also been reported that students and medical doctors are poorly trained on these technologies [323, 331–333]. One study indicated that, although a small cohort of UK medical students who received AI teaching felt more confident in working with AI in the future compared to students that did not receive teaching, a significant number of taught students still felt inadequately prepared [323]. In order to take full advantage of these technologies, scholars seem to agree that medical school training on AI should be expanded and improved [321, 323, 324, 328, 329].

Regarding the impact of AI on career choice and reputation, it was reported that AI has an impact in the career intentions of students with respect to radiology [323], but radiologists would still choose this specialty if given that choice [324]. These specialists have, however, revealed concerns that AI might diminish their professional reputation [327]. Contrary to the perceptions of the general public that AI will completely or partially replace human doctors [334], medical students and doctors in general are not concerned about job replacement [181, 321, 327, 329, 335].

Another important issue related to medical practice and career is liability. In a study in which pathologists were surveyed, it was reported that, with respect to medico-legal responsibility for diagnostic errors made by a human/AI combination, opinions were split between those who believed that the platform vendor and pathologist should be held equally liable, and others who believed responsibility remains primarily that of the human, with only a minority reporting that the platform vendor should primarily be liable [328].

Clearly, the ethics surrounding implementation of AI in healthcare goes beyond issues related to medical practice and career. Health AI gives rise to higher level ethical issues such as Autonomy, Fairness, or Privacy [152, 155] but, with the exception of fairness, these issues have received less attention in the scientific literature. Fairness concerns related to racial and gender bias in AI-powered medical applications have to do with the fact that AI algorithms are trained on predominantly male white patient data. Concerns have been raised both in popular and scientific literature about these algorithms perpetuating and amplifying existing bias and inequalities in healthcare [336–340]. It has been cautioned that medical data needs to be critically appraised in order to avoid such bias [336].

In this empirical study we surveyed medical doctors on a wider scope of ethical issues about AI in Healthcare. We addressed Privacy, Fairness, Accountability, Transparency, Safety, Human Oversight, Explainability, Future of Employment, Responsible Research Funding, Education about AI, Human Autonomy, Certification of AI products, Ethical Design. The diverse array of Health AI ethics surveyed in this empirical study allows us to discern the views and moral opinions of medical doctors about the implementation of AI in healthcare.

4.3. METHODS

4.3.1. OVERVIEW

In this research we used q-methodology, a systematic empirical approach to identify possibly conflicting perspectives of (stakeholder) individuals about a particular topic [341–345]. The core premise in q-methodology is that subjectivity is always self-referent, i.e. only the individual can measure his or her subjectivity, relational, i.e. the meaning of a statement is derived from its relation to other statements, and it can be demonstrated to have structure and form [343]. This method is therefore considered adequate for our purpose of systematically discerning and studying the subjective views of medical doctors about Health AI.

Q-methodology requires participants to sort a pre-defined set of items according to a subjective notion of agreement/disagreement. In this study medical doctors were invited to sort a set of statements retrieved from popular and scientific literature capturing key ethical issues about Health AI in a bell shaped distribution ranging from -5 to +5 and to provide provide additional comments about the statements they ranked highest (+5) and lowest (-5) [1]. Using statistical techniques, coherent clusters are formed which present particular perspectives into the ethics of Health AI. We interpret these perspectives and discuss in what ways they relate to and differ from one another.

There are great advantages in using q-methodology when compared to other exploratory research methods, such as interviews, focus groups, and surveys. Unlike interviews, q-studies provide numerical results to support subjective perspectives about a particular topic thus combining quantitative and qualitative approaches [346]. Moreover, because participants in q-studies sort items individually, these studies are less affected by dominance effects, which are observed in other research methods administered in groups, such as focus groups [346]. And unlike standard surveys, in which the opinions of participants about each topic are extracted separately, q-studies require participants to consider such topics simultaneously thus uncovering latent connections and allowing for more nuanced and sophisticated opinions [346, 347].

For the purposes of our study, which we recall is to reveal the diverse views about the ethics of Health AI, we also considered q-methodology to be a more suitable research method when compared to the Delphi method [348]. The latter is typically used for expert consultation and in that sense it is similar to our q-methodological study, in which we survey medical doctors about Health AI. However, the focus in the Delphi method is on reaching convergence (reducing heterogeneity) among experts about certain uncertain outcomes, whereas the q-method focuses on revealing the heterogeneity among stakeholders or experts.

This study followed the typical four phase sequence in q-methodological studies comprising (i) *definition of the concourse of communication*; (ii) *development of the set of statements (Q-set)*; (iii) *selection of participants (P-set)*; and (iv) *analysis and interpretation*. Below we provide further details about each one of these phases in this particular study.

[1] This study received ethics approval from the Human Research Committee of Delft University of Technology (letter of approval 1156).

4.3.2. CONCOURSE OF COMMUNICATION

The concourse of communication is a corpus of opinions related to a particular topic [349]. Such opinions can be gathered through direct sources, such as interviews and nominal group technique, or indirect sources, such as articles, discussion boards, and blogs. In this study we used quite varied indirect sources, including scientific publications as well as publications issued by popular science outlets, professional associations, consulting companies, and also blogs.

We reviewed scientific and gray literature on Health AI using combinations of keywords "Artificial Intelligence", "Machine Learning", and "Augmented Intelligence" along with connector "AND" and keywords "Healthcare", "Medicine", "physicians", and "medical doctors" in Google, Google Scholar, and Web of Science.

Initially we selected 353 statements for our concourse of communication and subsequently we assigned these statements to fifteen clusters, using a list of ethical issues compiled from 22 major guidelines of AI ethics as a guidance tool [152]. Each cluster of statements was associated with a particular AI ethical issue from such list, namely Privacy; Fairness; Accountability; Transparency; Safety and Cybersecurity; Human Oversight; Explainability; Future of Employment; Responsible Research Funding; Education about AI; Human Autonomy; Certification of AI products; Ethical Design; Health ppecific deliberations; and one additional cluster was added concerning AI in the Covid-19 Pandemic.

Organizing the concourse of communication in clusters that map onto overarching AI ethical issues facilitated the definition of the q-set, since the statements in this set should reflect the entire space of ethical issues identified in the concourse. It should be remarked, however, that the list of ethical issues used in this research as a reference tool reflects a particular deontological-based approach to Ethics. Other potential relevant ethical approaches and principles related, for instance, to informed consent and risk acceptance are therefore not included in such list [247]. Future research may identify and further explore additional ethical issues and values about Health AI.

4.3.3. SET OF STATEMENTS (Q-SET)

The q-set is a comprehensive yet manageable subset of the concourse of communication. We analyzed each statement in the clusters defined within the concourse in order to select the relevant items for a structured, comprehensive, and balanced set of statements. This selection was guided by three main considerations, namely, (i) accounting for a broad scope of positions put forward in the AI Health popular and scientific literature; (ii) favoring clarity; and (iii) avoiding redundancy. Using this method of obtaining the concourse, we have aimed for a maximum of objectivity and neutrality

The final q-set features 40 statements. Minor edits were made to these statements in order to ensure neutrality and also to meet the number of characters allowed by FlashQ [350], the software tool that was used in this study for administering the survey. The size of the set is at par with current q-methodology practices [343].

The landscape of statements in the q-set with respect to the pre-defined ethics clusters is composed of Privacy (statements 1-4); Fairness (5-8); Accountability (9-10,40); Transparency (11); Safety and Cybersecurity (12-13,39); Human Oversight (18); Explainability (15-17); Future of Employment (19-20, 22); Responsible Research Funding (23-24); Educa-

tion About AI (25,34); Human Autonomy (18); Certification of AI products (29-30); Ethical Design (31-33); Health specific deliberations (14,21,26-27,36-38); and AI in the Covid-19 pandemic (28,35). The final set is listed below.

1. Privacy should not be the highest priority in AI-based Healthcare.
2. Confidentiality should not constrain the implementation of AI in Healthcare.
3. Without clear rules about data usage, storage, and anonymization, AI should never be used in Healthcare.
4. Confidentiality, as defined today, has little use in a future where Healthcare relies heavily in AI.
5. AI is more likely to resolve rather than amplify inequalities in healthcare.
6. Improving equity and inclusion should be the top priority when developing and deploying AI in healthcare.
7. AI will increase discrimination based on predicted future medical problems.
8. We should be conservative in promoting AI in healthcare because of the unresolved ethical issues.
9. AI developers must be bound by medical ethics.
10. For the sake of technology advancement AI companies should not be liable for medical errors.
11. AI medical tools should only be used if clinicians understand how AI decisions are made.
12. There is high risk for monopolistic behavior by private AI companies in the domain of Healthcare.
13. It is undesirable that big companies enter the health care space because they know little about Medicine.
14. The patient-physician relationship will change dramatically once AI is fully deployed in health systems.
15. Health professionals do not need to know how AI medical tools work but rather if they are reliable.
16. Health professionals have always trusted black boxes (e.g. MRI) and it will not be different with AI.
17. Appropriate informed consent is not possible if the medical doctor cannot explain to the patient how the AI medical device works.
18. AI will decrease the autonomy and authority of medical doctors.
19. AI will not replace doctors, but doctors who use AI will replace doctors who do not.
20. If AI tools work well, Hospitals should save money by hiring less highly skilled practitioners.

21. AI will worsen problems in healthcare such as overtesting, overdiagnosis, and overtreatment.
22. Automation may work well in factories, but not in Hospitals.
23. AI-based medical products won't be able to match the hype.
24. All the funding allocated for AI is worthwhile if it can take over bureaucratic shores, such as note-taking, coding, and pattern-finding.
25. Doctors are not interested in learning about AI and Computer Science.
26. In the medical field it is problematic that machines lack contextual knowledge and ability to read social clues.
27. It would be unethical not to use AI tools if they provide better decisions than medical doctors.
28. AI has already played a vital role in the COVID-19 pandemic.
29. The mantra of the tech industry "fail fast and fix it later" is putting patients at risk and regulators are not doing enough to keep consumers safe.
30. AI healthcare products must be tested in randomized clinical trials, which is the strongest source of medical evidence.
31. Because AI systems are designed mainly to increase profit, in the future health systems will have more resources and provide better care.
32. Healthcare AI technology must be aligned with bioethical principles.
33. Medical doctors must participate in the design process of AI for Healthcare.
34. Clinicians lack the time to learn how to use complex AI-based medical devices.
35. AI enhances medical decision making in situations of care rationing.
36. AI will allow providers, clinicians, and staff, to focus on more top-of-license skill sets and activities.
37. Most areas of healthcare can benefit from AI.
38. It is not very difficult to operationalize clinical practice for a machine.
39. Medicine should never rely on AI because such computer systems are vulnerable to cybersecurity threats.
40. If a medical doctor makes a mistake as a result of the advice from an AI tool, he/she should be considered liable.

4.3.4. PARTICIPANTS (P-SET)

In the recruitment of participants for this study three different approaches were used, which entailed reaching out to: (i) Hospital departments (through phone and subsequently by e-mail); (ii) medical doctors who are personal acquaintances; and (iii) medical doctors who are not personal acquaintances (through email addresses made available in publications found in Google Scholar related to various medical fields). Provided that approach (iii)

proved to be much more successful, eventually the other strategies were dropped and we focused mainly in reaching out medical doctors through publications they had recently authored.

Rather than focusing on a particular medical field, in which the practitioners may share similar thoughts about AI, we aimed at including a diverse mix of specializations which would allow us to have a wider breadth of viewpoints in the data.

In order to select publications from different medical fields, we used keywords "Surgery", "Anesthesiology", "OBGYN", "Gynecology", "Ophtalmology", "Intensive Medicine", "Neurology", "Family Medicine", "Primary Care", "Radiology", "Nuclear Medicine", "Neuroradiology", "Pathology", "Rheumatology", "Oncology", "Dermatology" along with connector "AND" and keywords "Netherlands", "Portugal", and "United States" in Google Scholar. Subsequently, through snowballing techniques, additional relevant articles and scholars were identified.

Provided that this study aimed at surveying medical doctors, when the corresponding author of a scientific article was not identified as a MD in the publication, we did additional searches in Google to confirm if the scholar was indeed a medical doctor. Each participant was therefore contacted through the e-mails made publicly available in the scientific publications, in the capacity of being an author or co-author of a particular publication as well as a medical doctor.

The final set of participants in this study comprised medical doctors (residents and specialists) from thirteen different specialities including medical specialties (Family Medicine, Rheumatology, Dermatology, Intensive Medicine, Oncology, Neurology), surgical specialities (Surgery, Ophthalmology, OBGYN, Anesthesiology, Rehabilitation Medicine, Neurology), and diagnosis specialties (Pathology, Radiology/ Nuclear Medicine/ Neuroradiology) based in The Netherlands, Portugal, and U.S. Further details are found in Table 4.1 below.

A total of 77 participants successfully completed the survey, which is an adequate number for a q-methodological study featuring a q-set of 40 items [341, 343]. Indeed because q-methodology aims just at establishing the existence of particular viewpoints, large numbers of participants are not required. Moreover, q-studies do not require a rigorously representative sample but rather a population sample that contains participants with relevant viewpoints on the matter.

We are confident that the p-set in this study includes scholars with relevant viewpoints on AI Health. However, we acknowledge that, by targeting medical doctors who had recently published scientific articles, the set of participants is mainly composed of practitioners who are involved in research and/or academic activities. We may therefore have failed to represent other perspectives from practitioners that are less involved in research.

In this context, it should also be clarified that a q-study typically makes no claim that the relative sizes of the perspectives (in terms of the number of respondents that adhere to them) reflect the population distribution. In keeping with the notion that q-methodology is an exploratory rather than a confirmatory technique, and acknowledging the way in which the sample was obtained, we will refrain from drawing any quantitative conclusions about sizes of perspectives and differences between countries and specializations. Follow up confirmatory research (e.g. aiming at establishing minority versus majority views) should be based on representative samples.

Table 4.1: Participants (p-set).

Specialization	Portugal	Netherlands	U.S.
Surgery	1	3	4
Anesthesiology	1	2	6
OBGYN	3	2	0
Ophtalmology	3	1	2
Rehabilitation Medicine	1	1	1
Intensive Medicine	1	2	0
Neurology	1	2	1
Family Medicine	2	1	2
Radiology/Nuclear Medicine/Neuroradiology	4	3	0
Pathology	1	6	0
Rheumatology	8	3	1
Oncology	1	1	2
Dermatology	1	3	0

4.3.5. SURVEY COLLECTION TOOL

The data was collected through the html version of FlashQ [2], a software that allows online q-sorting. The distribution was coded as a 11-point distribution resembling a normal distribution [-5, +5] with two cells placed under each tail (-5 and +5), three cells under both -4 and 4, three cells under both -3 and 3, four cells under both -2 and 2, five cells under both -1 and 1, and six cells under 0 (Figure 4.1). Participants were asked to arrange the 40 statements according to a subjective notion of disagreement/agreement and subsequently were asked to provide further comments on the statements they ranked -5 and +5. Each particular arrangement of the statements in the forced bell-shaped distribution is called a q-sort so in this study the collected data consisted of 77 q-sorts.

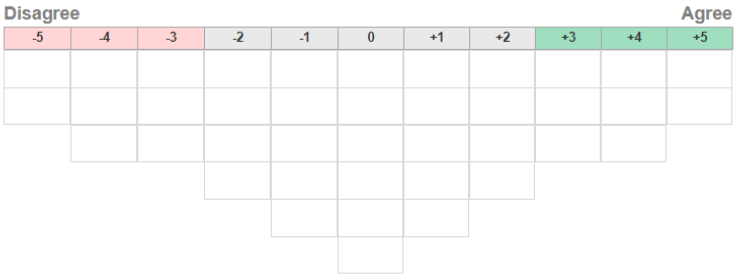


Figure 4.1: Sorting grid: participants sort the statements in the bell-shaped distribution.

[2] Q-methodology software packages and resources are available https://qmethod.org/resources/software/. The html version of FlashQ used in this research is no longer available at http://www.hackert.biz/flashq. A html version of FlashQ can be currently found in https://shawnbanasick.github.io/ken-q-analysis/ [350].

4.3.6. ANALYSIS

As a derivation of factor analysis, q-methodology is a data reduction technique which aims to reduce a larger number of variables into fewer factors. Therefore, the analytic process of q-methodology relies on multivariate data-reduction techniques. In q-studies, data analysis entails three main steps: (i) factor extraction; (ii) factor rotation; and (iii) factor interpretation. In the analytic process (steps (i) and (ii)) we used PQMethod, a statistical program that accommodates the requirements of q-studies [3] [351].

The first step consists of extracting factors from previously collected q-sorts thus summarizing all individual responses into a few representative responses [346]. In this study, the factors were extracted through Principal Component Analysis (PCA), a linear reduction technique that reduces the dimensionality of the data while retaining most of the variation in the dataset, which is often used in exploratory data analysis [352]. The identification of orthogonal vectors (principal components) along which variation is maximal allows the reduction of data into a few components that represent the dominant patterns in the data [352, 353].

Table 4.2: Overview of factors.

	Factor 1	**Factor 2**	**Factor 3**	**Factor 4**
Defining sorts	15	17	6	9
Eigenvalues	17.77	11.55	8.47	10.78
Variance	17%	15%	11%	14%

[3] http://schmolck.org/qmethod/pqmanual.htm

Table 4.3: Four factor arrays where each array features the normalized scores [-5, +5] assigned to the statements in the q-set by participants who loaded significantly on the factor array.

Statement	Factor 1	Factor 2	Factor 3	Factor 4
1	0	-4	-5	-3
2	-1	-5	-3	-1
3	2	4	-3	2
4	-1	-4	-2	-4
5	1	-3	2	0
6	0	2	-1	1
7	0	0	-3	1
8	-1	1	1	0
9	4	5	4	4
10	-2	-5	-4	-2
11	-2	2	3	5
12	1	4	-1	2
13	-2	1	-5	-1
14	-2	-2	1	0
15	2	-1	1	-4
16	1	-2	-1	-2
17	-3	1	2	-2
18	-5	-1	0	-1
19	4	0	0	1
20	-5	-4	-1	1
21	-4	2	-2	-2
22	-4	-1	-4	-3
23	-1	1	-2	0
24	3	2	2	2
25	-4	-2	0	-5
26	0	3	4	4
27	3	0	1	2
28	1	-2	-1	0
29	0	3	1	0
30	3	5	5	-1
31	-1	-3	0	-1
32	4	3	5	4
33	5	4	3	5
34	0	0	0	-4
35	2	-1	3	1
36	5	0	2	3
37	2	1	4	3
38	-3	-3	-2	-5
39	-3	-1	-4	-3
40	1	0	0	3

The extracted factors were subsequently rotated in order to position each factor so that its viewpoint closely approximates the viewpoint of a particular group of q-sorts. For the rotation of the factors, we used Varimax, an orthogonal rotation of the factor axes that maximizes the variance of each factor loading by making high loadings higher and low loadings lower. Q-sorts that load high on one factor will load low on another, thus maximizing the distinction and differentiation of subject positions while minimizing the correlation among factors [354]. Upon rotating different numbers of factors and comparing the distributions of (automatically flagged) defining sorts among factors, a decision was made to rotate four factors (Table 4.2). This solution features the highest yet interpretable

number of factors in which every factor has at least three defining sorts.

Each factor is characterized by a factor array featuring 40 scores (one score per statement), which is a single q-sort configured to represent the viewpoint of the factor. Given that factors have different numbers of defining sorts, each score in the factor array is a standardized (z) score to allow cross-factor comparison [355]. The factor arrays of each factor are shown in Table 4.3.

Finally, the last step entailed analyzing and interpreting the factor arrays of the four perspectives in order to understand the key features of each perspective. For this purpose we used the crib sheet method [343]. By looking at the factor arrays, for each perspective we composed four basic categories: (i) items with highest ranking in the factor array; (ii) items with lowest ranking in the factor array; (iii) items ranked higher in factor i than in any of the other factors (by 2 or more units); (iv) items ranked lower in factor i than in any of the other factors (by 2 or more units). It is noted that the interpretation of the factors made by the authors is (inherently) subjective. It is possible that different people may arrive at different interpretations of the four factors based on the same factor scores. Yet, given that these interpretations are constrained by the factor scores, we would expect that other researchers would arrive at similar interpretations.

4.4. RESULTS: PERSPECTIVES & INTERPRETATIONS

Four different perspectives about Health AI were identified in this study (Table 4.4). The core characteristics of each perspective are derived from the statements ranked -5 and +5 [$(N :| 5 |)$ where N is the number of the statement and $(| 5 |)$ may be either -5 or +5] as well as the statements ranked highest or lowest compared to the arrays of the other perspectives [$(N :| P_i |)$ where N is the number of the statement, P_i is the perspective with $i \in [1, 4]$, and $| P_i |$ may either be $-P_i$ or $+P_i$ depending if the statement is ranked lowest or highest than in the arrays of other perspectives]. For the purpose of further illustrating each perspective, we also included statements written by participants associated with the defining sorts of each perspective, about the statements they ranked highest and lowest [4].

Table 4.4: Four perspectives about Health AI.

P1	AI is a helpful tool: Let physicians do what they were trained for
P2	Rules & regulations are crucial: Private companies only think about money
P3	Ethics is enough: Private companies can be trusted
P4	Explainable AI tools: Learning is necessary and inevitable

4.4.1. PERSPECTIVE 1
AI IS A HELPFUL TOOL: LET PHYSICIANS DO WHAT THEY WERE TRAINED FOR

In this perspective there is an overall positive outlook about the implementation of AI technology in healthcare. AI is regarded as a helpful tool that will allow doctors to have the time to focus on top-of-license skill sets and activities (36:+5).

[4] All comments written by participants are available in Appendix A.

Underlining this position, one participant wrote *That is the main aim! To let physicians do what they were trained for - medicine - and alleviate many of the potentially automatic and time-consuming processes they have to daily face.* Another participant noted that AI means *less time needed for boring work means more time for challenging work.* And yet another participant reflected on his early days in the medical field to make a point about the positive aspects of automation *Much like automation for lab tests, AI will free up the providers' hands and mind to focus on higher order issues. As an intern, I had to spin my own hematocrits at night. I do not miss that at all!*

Traditional arguments raised against Health AI are understated in this perspective. It is not problematic that AI is a black box technology since health professionals have been using other black box technologies, such as MRI (16:1). And there is also a neutrality about AI's lack of contextual knowledge and ability to read social clues (26:0). Along these lines, one participant noted that *The role of a skilled physician is to take into consideration what a machine / AI tells him and make the correct connection with clinical reality.* Moreover, this perspective does not subscribe to the thought that AI will worsen problems in healthcare such as overtesting, overdiagnosis, and overtreatment (21:-4). About the role of AI in Hospitals, one participant wrote that *AI is going to play a pivotal role in stratification, thereby assigning patients into low-risk and high-risk groups or patients responding to a certain treatment or patients not responding to it. This will prevent testing or treating patients in whom it is deemed not efficient.*

Despite the positive outlook about AI, this perspective emphasizes that medical doctors must remain in charge not only in the medical decision process (18:-5) but also by participating in the technology design process (33:+5) (*AI will only help the physicians resolving their clinical doubts, but the last decision should never be given by AI*; *I see AI as an additional tool, not as something that will replace MDs or decrease autonomy and authority. A MD will always have the final verdict*; *AI designers know the technology, but need MDs to design relevant products*; *The goal of a certain AI tool has to be defined together with the medical doctors to ensure clinical relevance*; *As doctors lack informatics skills, engineers lack medical knowledge and hospital needs, therefore medical Doctors are key in the design of AI.*)

Looking forward to the future of medical employment, according to this perspective AI-based hospitals should not save money by employing less skilled doctors (20:-5) (*With AI tools working well, Medicine will advance to a more precise act, with decision based on multidisciplinary team opinion, so highly skilled practioners will be most needed*) but it is considered that even though AI will not replace doctors, doctors who use AI will replace doctors who do not (19 -4) *As in any area of technical progress, AI is a tool that will be embraced by those at the cutting edge. Those who don't – like surgeons who never mastered laparoscopy – will find their scope of practice diminishing.*

4.4.2. PERSPECTIVE 2 RULES & REGULATIONS ARE CRUCIAL: PRIVATE COMPANIES ONLY THINK ABOUT MONEY

In the second perspective identified in this study, there is a clear negative outlook about AI technology (21:2 ; 31 -3) and a clear distrust in private health companies. There is a sentiment that the tech industry is not well aligned with core healthcare values (*Private companies only think about MONEY*), has little knowledge about Medicine (13:1), and

poses a risk for monopolistic behavior (12:4).

About the risk for monopolistic behavior, there is a great concern about the implications of these big companies owning medical data. Privacy is heralded as a core ethical value in the medical field also in a future where healthcare relies heavily on AI (2:-5; 4:-4). In particular, there are concerns about the implications on the patient physician relationship. A participant noted that *when a patient is worried that confidentiality is breached towards a tech company or insurance company, they may not provide full information or avoid treatment leading to a more profound disease.* There are also concerns about the power that would come from owning such data. A participant cautioned that *Healthcare AI companies would have too much power if this information was not anonymous. They would sell information about specific people to drug companies, to hiring companies, to insurance companies .....*

Because in this perspective private companies are not to be trusted, there is a strong emphasis in rules and regulations to keep these corporations in check. It is considered that the tech mantra "fail fast and fix later" is putting patients at risk and that regulators are not doing enough to keep consumers safe (29:3). Therefore, even though Health AI technologies and its developers should be bound by core medical ethics (9:5), clear rules about liability, data, and product certification must also be in place.

Moreover, it is perceived that technology companies must be liable for medical errors even if such liability hampers technological advancement (10:-5). A participant elaborated that *profit demands risk and the companies must bear that risk. Developers must be accountable if they wish to enter the demanding arena of care.* These technologies should only be used in healthcare once clear rules and regulations about data usage, storage, and anonymization, are in place in AI (3:4).

Along the same lines, it is also emphasized that AI health products must be tested in randomized clinical trials, which is the strongest source of medical evidence (30: +5). One participant illustrated this point quite clearly by stating that *since the diagnostic and treatment AI tools affect directly the patients health, they should be held against the highest standards as usual in medicine for new diagnostic and treatment strategies. I do not see why this should be different for AI then for new " conventional diagnostic tests" or drugs.*

4.4.3. PERSPECTIVE 3 ETHICS IS ENOUGH: PRIVATE COMPANIES CAN BE TRUSTED

The most striking feature about this perspective is the overall positive outlook about AI companies. According to this view, it is not undesirable that these companies start operating in the health space (13:-5). Moreover, there are no major concerns about the risk for monopolistic behavior (12:-1).

A potential explanation about this positive account on tech is the perception that current health systems already rely heavily on technology and tech companies. As one participant noted, about the potential of automation in Hospitals, *this is ridiculous. Automation already works in Hospitals.*

Rather than dwelling on rules and regulations (3:-3), ethics in itself is enough to ground the private sector. AI technology must be aligned with bioethical principles (32:+5) such as Privacy (1:-3; 2:-3), which should remain a core medical value (4:-2). Despite the trust

in tech companies, also in this perspective the need for testing of AI health products is emphasized (30:5).

According to this perspective, AI will not increase discrimination based on predicted future medical problems (7:-3) and therefore improving equity and inclusion is not mandated to be the top priority when developing and deploying AI in healthcare (6:-1).

4.4.4. PERSPECTIVE 4: EXPLAINABLE AI TOOLS: LEARNING IS NECESSARY AND INEVITABLE

Explainability is a key value in this perspective. In order to reap the benefits of AI, medical doctors must understand and lead the AI technological progress. A participant wrote *AI should never be a "black box". Doctors should be able to explain the results from AI tools with reasoning.* Not only health professionals need to know how AI medical tools work (15:-5), but in fact such tools should only be used if clinicians understand how AI decisions are made (11:5). Along these lines, one participant remarked that *the adoption of AI will be improved if doctors do understand the 'black box'.*

According to this perspective, doctors are interested in learning about AI and Computer Science (25:+5) and have the time to learn how to use complex AI-based medical devices (3:-4). One participant noted that *overall, the intellect of doctors is underestimated and under evaluated by technicians*; and another highlighted that *doctors cannot work without computers and use them daily for registration. Learning is necessary and inevitable.*

It is considered problematic that machines lack contextual knowledge and ability to read social clues (26:4) and it is difficult to operationalize clinical practice for a machine (38:-5). Therefore, medical doctors must participate in the design process of AI for Healthcare (33:5). Accordingly, a participant wrote that *AI is here to stay (I think), and medical doctors are the most suited to adjust and improve the various algorithms etc. that are currently being designed.*

4.5. DISCUSSION

The perspectives identified in this study reveal diverse and often contradictory viewpoints about Health AI. Understanding these underlying values and tensions is important for operationalizing the ethical issues associated with the implementation of AI technologies in healthcare. Ultimately, such operationalization is expected to lead to more meaningful debates and policies towards an ethically aligned deployment of Health AI.

Our study offers a systematic analysis of the perspectives of medical doctors about Health AI. It is possible to observe elements of the four reported perspectives in the current literature. Several articles have reported findings that trace back to P1 (AI is a helpful tool: Let physicians do what they were trained for) with respect to the positive attitudes about the use of AI as a supportive technology [321, 322, 327–330]. The need for AI medical school training to be expanded and improved is also well addressed in the literature [321, 323, 324, 328, 329] and relates to P4 (Explainable AI tools: Learning is necessary and inevitable), which contends that doctors must understand AI. Moreover, none of the perspectives identified in this study reveal concerns about job replacement, which is a finding aligned with studies reported in the literature [181, 321, 327, 329, 335]. The strength of our work

lies on the fact that, because participants were surveyed on a wide range of ethical issues related to Health AI, the perspectives that have emerged provide a more comprehensive picture of the moral views of practitioners.

Each perspective provides insights about the AI Health ethics outlook and also about particular ethical issues, such as Fairness, Explainability, and Ethical Design, which need to be accounted for on the implementation processes of AI technologies in healthcare.

About the outlook on AI Health Ethics, P1 and P2 (Rules & Regulations are crucial: Private companies only think about money) represent somewhat conventional views about AI technology, which contrast with P3 (Ethics is enough: Private companies can be trusted). The perception that AI-based medical tools will improve efficiency in the clinical setting builds on decades of successful development of sophisticated medical technologies. In general, this perspective (P1) is aligned with the narratives presented by tech companies, which tend to focus on the benefits of technology and automation to take over repetitive tasks. Similar accounts are presented by developers of other AI-based technologies, such as the autonomous vehicle [167].

By projecting AI-based medical technologies as yet another type of medical tool, higher order conversations about ethical issues associated with this technology, such as Fairness or Human Autonomy, are to some degree avoided. However, because of the disruptive potential of these technologies, further thoughts about ethics are required. Tech companies and developers should indeed acknowledge the singularity of AI and ensure that the design process is to be guided by ethical considerations.

When the emphasis is not just on the benefits of AI technology, regulation is often seen as the solution to ensure the safety of consumers. Medical doctors who are hesitant about AI-based medical tools, consider that rules and regulations are a crucial element in the transition for AI-based healthcare. This perspective is well aligned with the tradition in the healthcare and medicine fields, which are notoriously heavily regulated [356]. The regulation of AI-based medical devices (also known as Artificial Intelligence/Machine Learning-Based Software as a Medical Device) has unique challenges and is known to be lagging behind the technology [357].

In contrast, a less conventional perspective with respect to the ethics of AI Health reported in this study (P3: *Ethics is enough: Private companies can be trusted*) considers that heavy regulation of tech companies is not needed. By refusing to demonize AI tech companies, this perspective provides a somewhat unusual positive outlook about these health stakeholders. It builds on the idea that ethical awareness is enough for corporations to be trusted. Recently it has been argued that AI corporations should indeed promote virtue ethics, rather than the traditional deontology-infused guidelines, as an effective form of ensuring ethical behavior in corporations [152]. While completely forgoing regulations may be unrealistic, private tech companies should indeed internalize that to be accepted as an ethical player in the health space while also striving for profits, must promote ethical environments and practices.

About particular ethical issues, our study unraveled tensions and contradictory viewpoints that should be accounted for in future debates, namely with respect to Fairness, Explainability and Ethical Design.

Fairness, non-discrimination, and justice relate to reasonableness and impartiality of

actions. AI-based technologies are designed and produced by humans and rely extensively on data thus being exposed to errors, ill judgments, and prejudices which can enter into the innovation lifecycle and create biases [87, 358]. There are several concerns about biased or discriminatory outcomes in the context of Health AI. A biased medical device operates in such a manner that produces disadvantages to certain demographic groups and influences health inequality [359]. Different types of bias are associated with medical AI-powered devices, namely physical bias (design of the medical device disadvantages certain demographic groups based on physical traits such as skin color), computational bias (training datasets that serve as inputs in medical device are not representative of population), and interpretation bias (medical device is subject to biased inference of readings) [359].

Our study indicates that there are more concerns about these matters in P2 and less in P3, which is not surprising given the remaining features of these perspectives. However, in general, all perspectives are quite neutral when it comes to fairness and discrimination in Health AI. This neutrality could well be a short-coming of this study, namely related the selected statements about Fairness, but could also signal that medical doctors are ill-informed about these issues or they just do not consider them as relevant or pressing. Recently some concrete cases about unfair medical devices have been reported in the literature (for instance it has been reported that pulse oximeters are not as accurate in measuring blood oxygenation in Black patients [359]), but it could be the case that these matters remain largely abstract for the majority of practitioners. Future empirical research should further explore the views of practitioners on Fairness and bias issues.

Another important issue associated with the ethics of Health AI is Explainability. An explainable model provides *interpretable* (description of a system in a way that can be understood by humans) and *complete* (accurate description of the operation of a system) information about the system [360]. The challenge of explainability is therefore to reach both interpretability and completeness, given that accurate explanations are not easily interpretable and the latter often lack predictive power [360]. AI-powered medical technologies rely on complex algorithms which are not easily interpretable, thus known as black-boxes.

Our study revealed contrasting viewpoints with respect to explainable Health AI technologies. According to P1, the lack of explainability in AI-powered devices is not problematic, since health professionals have been using other complex technologies, such as MRI, which also resemble black-boxes. In contrast, P4 considers that Explainability is a key value and that in order to reap the benefits of AI, medical doctors must understand the intricacies of AI powered medical devices.

The comparison with MRI is often called for in this literature but it is not widely accepted. Indeed the MRI is a complex medical technology and practitioners are not expected to know the underlying physics and math of this technology. However, the algorithms that operate these systems are indeed explainable and understood by developers. Differently, the explainability challenge associated with AI is not contingent to medical practitioners but also to developers in general. Lack of knowledge about the decision rules that sustain a certain outcome is especially problematic in the healthcare setting. As one participant remarked, *A good health careprofessional will never blindly rely on any single measure without the story of the patient.*

The contrasting views reported in our study support the need for further empirical research

in order to determine whether practitioners who share different perspectives with respect to Explainability would interact differently with the same algorithm [361].

Both values of Fairness and Explainability explored above should be accounted for in the Ethical Design of Health AI. Our study shows quite clearly that, regardless of the positive or negative industry outlook, all perspectives consider that medical doctors must participate in the design process of AI health technologies.

By further exploring the comments of participants, it seems that medical doctors even consider that the success and clinical relevance of AI Health depends on the involvement of practitioners in design and development of the technology. The reasons advanced by participants go beyond medical knowledge (*As doctors lack informatics skills, engineers lack medical knowledge and hospital needs, therefore Medical Doctors are key in the design of AI.*), and include also clinical reasoning (*Without knowing how a clinical thinks, than AI would not be a useful tool*), and the societal role of medical practitioners (*MDs are trained and dedicated in ethical and societal decision making. They are natural bridge builders between a complex medical/technical reality and the personal space of an individual patient. Crossing this bridge is fundamental for any novel development to have a reason of existence. Without MDs involved, the major stakeholders of the development are neglected*).

The development of AI Health technology should be a multi-disciplinary effort. Whether practitioners will act as advisors or, as cautioned in P4, will take a more prominent role in the tech development is yet to be seen. Further research should explore models of technology development that are able to integrate the elements mentioned above in the design process.

Our exploratory study revealed four perspectives about Health AI, which we expect may help to shape future debates as well as ethical design processes. There are contrasting views about the Ethics of Health AI in general but also about particular ethical issues such as Explainability. It is observed that medical doctors are more concerned about the role of large companies in healthcare and less aware or concerned about higher level and often abstract issues such as Fairness, bias, and health inequalities.

There are important limitations in this study. The first limitation is related to the filtering process, in which the set of 353 statements retrieved from scientific and popular literature was reduced to 40 statements. This process was carried out by the authors, which do not have medical training. We acknowledge that having the input of a medical doctor in such filtering process would add value to this study by allowing us to have a better understanding of the relevance and knowledge of medical practitioners about the matters captured in the statements. Another limitation is related to the recruitment of participants. As mentioned earlier, the vast majority of participants in this study was recruited through recent scientific publications. We may therefore have failed to capture perspectives of practitioners that are less involved in academic and research activities. By recruiting medical doctors from three Western countries we also failed to include in our p-set medical doctors from developing countries which may have contributed with additional perspectives about Health AI. Finally, while in this study we have focused on perspectives regarding AI-applications in Healthcare in general, there is a need to expand the literature and look into particular domains and tasks of Health AI.

4.6. CONCLUSION

For AI to meet its potential in the complex Healthcare space, ethics needs to be taken into account. In this empirical study we surveyed medical doctors based in The Netherlands, Portugal, and U.S. on a wide scope of ethical issues about Health AI. This survey allowed us to discern different perspectives about the ethics surrounding the deployment of Health AI.

We identified four main perspectives: P1: *AI is a helpful tool: Let physicians do what they were trained for*; P2: *Rules & Regulations are crucial: Private companies only think about money!*; P3: *Ethics is enough: Private companies can be trusted*; and P4: *Explainable AI tools: Learning is necessary and inevitable!*.

Each perspective provides valuable insights about ethical issues that should be operationalized and accounted for in the design and development of these technologies. Our study reveals contrasting viewpoints about the ethics associated with Health AI. It is also observed that medical doctors are mostly concerned about the role of large companies in healthcare and less aware or concerned about higher level issues such as fairness, bias, and health inequalities. Regardless of the positive and negative industry outlook, our study revealed that medical doctors consider that they must participate in the design process. These findings are useful starting points for a fruitful discussion between medical professionals, industry stakeholders, and policy-makers.

Given the exploratory nature of this research, there is ample opportunity for confirmatory research directions and to explore how to translate these perspectives into actionable insights and design models for the different health stakeholders.

III

AI AS ARTIFICIAL MORAL AGENT

5

PERSPECTIVES ABOUT ARTIFICIAL MORAL AGENTS

A. Martinho, A. Poulsen, M. Kroesen, and C. Chorus, *Perspectives about artificial moral agents*, AI and Ethics **1**, 477 (2021).

ABSTRACT

The pursuit of Artificial Moral Agents is complicated. Disputes about the development, design, moral agency, and future projections for these systems have been reported in the literature. This study aims to provide empirical insights about those disputes by surveying (AI) Ethics scholars. Here we show the wide breadth of viewpoints and approaches to artificial morality. Five main perspectives about AMAs were empirically identified and discussed:(i) Machine Ethics: The Way Forward; (ii) Ethical Verification: Safe & Sufficient; (iii) Morally Uncertain Machines: Human Values to Avoid Moral Dystopia; (iv) Human Exceptionalism: Machines Cannot Moralize; (v) Machine Objectivism: Machines as Superior Moral Agents. The findings of this study have implications for the Machine Ethics project, which is reported to be primarily perceived as either the best way forward to realize ethical machines or as futuristic and lacking practical application of moral considerations. The diverse perspectives about artificial morality identified in this study unravel the need for grounding and practicality in Machine Ethics. If Machine Ethics is to be accepted as an applied ethic, it needs to be shaped to the field in which it operates and not the other way around.

5.1. INTRODUCTION

The development of Artificial Moral Agents (AMAs), i.e. artificial systems displaying varying degrees of moral reasoning, is an open discussion within the realm of Artificial Intelligence (AI). Given the rapid progress and pervasiveness of AI in modern society, there have been debates about the prospects of equipping these increasingly autonomous agents with moral machinery [34, 182, 362–364]. The endeavor of developing such an AMA is central to the Machine Ethics project [132, 160] and it is quite controversial [183, 365, 366].

There is an array of existing and projected systems that qualify as AMAs [362]. Existing, em-

pirically evaluated AMAs include *GenEth*, a general ethical dilemma analyzer that utilizes inductive logic programming to learn new ethical principles in-situ [367], and *Vanderelst and Winfeld's consequentialist machine*, which relies on functional imagination simulations to predict moral consequences [368]. Theoretical AMAs include the *Virtuous AMA*, which aims to observe and replicate human moral behavior by having the AMA learn and build character over time as per virtue ethics theory [369], and *MoralDM*, which models and weighs known psychological findings about utilitarian and deontological modes of reasoning, based on ethicists decisions in moral dilemmas, to inform action or inaction in novel moral decisions [370]. Most of the controversies surrounding AMAs concern projected AI Systems that rank high on the Autonomy and Ethics spectrum [182].

The controversial AMA debate is marked by conceptual confusion, excess of taxonomy, and practical inertia [371]. Particularly, there is a poor perception on the views and agreements within the (AI) Ethics communities on fundamental matters associated with AMAs, such as whether these systems should even be developed [183, 372], how to develop them [373], if they would have moral agency [374], and their moral and societal role [362, 375–377]. Although ambiguity is expected when it comes to Morality, given the interdisciplinary nature and pressing relevance of the subject matter, it is crucial to strive for some clarity on these fundamental matters.

The aim of this exploratory research is to uncover the diversity of views within the (AI) Ethics research community about key disputes surrounding AMAs, thus bringing coherence and clarity to these debates and ultimately allowing more insightful research avenues and policy recommendations. Understanding different views and, where possible, reaching an agreement is a common endeavor in a debate. To realize this aim we used q-methodology, an exploratory and semi-quantitative research methodology that provides a clear and structured way to elicit subjective views on particular issues and categorizes these viewpoints into clusters of value positions [341, 346].

Five main perspectives about AMAs emerged from our data and were subsequently interpreted and discussed: (i) *Machine Ethics: The Way Forward*; (ii) *Ethical Verification: Safe & Sufficient*; (iii) *Morally Uncertain Machines: Human Values to Avoid Moral Dystopia*; (iv) *Human Exceptionalism: Machines Cannot Moralize*; (v) *Machine Objectivism: Machines as Superior Moral Agents*. These perspectives represent different views and categorize agreements and disagreements about AMA development, design, moral agency, and future prospects.

The study findings bring coherence and clarity to disputes surrounding AMAs by organizing, specifying, and making clear the broader perspectives about these artificial systems. A more informed debate can continue with disagreements disclosed and appreciated. Moreover, some baseline agreements on particular topics are worth pointing out. Going forward, shared research principles could be developed based on those agreements.

This article is organized as follows: in the second section the methods used in this empirical research are described; in the third section a background on the four key matters associated with AMAs surveyed here is provided; in the fourth section the results are presented, i.e. descriptions of the five perspectives identified in this study; in the fifth section the results are discussed; and finally the sixth section features the concluding remarks.

5.2. METHODOLOGY

5.2.1. OVERVIEW

The methodology used in this research is q-methodology, a systematic empirical approach derived from traditional factor analysis, to determine the subjective views of individuals about a particular topic [341–345]. Q-methodology aims to bring coherence to complex and controversial matters by reporting on the significance assigned by participants to those matters [341, 342]. It is therefore deemed adequate to bring coherence to the controversial ethical matters related to AMAs.

Participants in q-methodological studies are required to rank order a set of items (e.g. statements) relative to one another on a grid that typically follows a bell shaped distribution. Subsequently, they are offered the opportunity to provide additional comments about the items they ranked highest and lowest according to a subjective dimension of agreement/disagreement. This last feature is of particular importance in this study, as the surveyed scholars provided interesting and often thought provoking comments that enrich the discussion about AMAs.

The statistical operations take place not in the columns but in the rows of the data matrix. One implication of this inversion from traditional by-variable to by-person factor analysis is that participants become the variables. Each revealed factor therefore has the potential to identify groups of persons who share the same perspective about a particular topic [343].

The unique features of q-methodology offer great advantages when compared to other exploratory research methods, such as interviews, focus groups, and surveys. Q-studies provide numerical results to support subjective perspectives about a particular topic thus combining quantitative and qualitative approaches [346]. Unlike standard surveys, in which the opinions of participants about each topic are extracted separately, q-studies require participants to consider such topics simultaneously thus uncovering latent connections and allowing for more nuanced and sophisticated opinions [346, 347]. Q-methodology also offers some advantages in mitigating response bias. By requiring participants to sort a pre-defined set of items, these studies are less prone to response bias since the participants are required to explicitly engage with views they disagree with or may have never considered before. Moreover, because participants sort the items individually, q-studies are less affected by dominance effects, which are observed in other research methods administered in groups, such as focus groups [346].

This study followed the typical four phase sequence in q-methodological studies comprising (i) *definition of the concourse of communication*; (ii) *development of the set of statements (q-set)*; (iii) *selection of participants (p-set)*; and (iv) *analysis and interpretation*. Further details about each one of these phases in this particular study are provided below.

5.2.2. CONCOURSE OF COMMUNICATION

For the definition of the concourse of communication, we reviewed scientific and popular literature on AMAs. A keyword search using word combinations “Machine Ethics”, “Artificial Moral Agents”, “Ethical Agents”, “Ethical Artificial Intelligence”, “Moral Artificial Intelligence”, “Moral Machines”, and “Autonomous Vehicles AND Ethics” in Google, Google Scholar, Web of Science, and Scopus allowed us to identify 44 relevant scientific articles

from which we extracted 167 statements.

In addition to scientific literature, we also looked for relevant popular science publications. Through online searches on Google, we identified 17 articles in popular science outlets such as Scientific American, MIT Technology Review, or Philosophy Now and extracted 36 statements. As a result, the concourse of communication of this study features a total of 203 statements. These statements represent often controversial and thought provoking propositions about AMAs Ethics.

Although the literature on AMAs is particularly nuanced and rich, recurrent topics were clearly identified. We observed that most publications address issues related to the morality of the quest for developing AMAs, design strategies to equip artificial systems with morality, moral agency of advanced artificial systems, and projections about the future moral and societal role of these systems. We have therefore considered these themes central for this research.

Accordingly, we assigned the statements composing the concourse of communication to four different clusters reflecting the themes mentioned above: (i) *Development of AMAs*; (ii) *Design of AMAs*; (iii) *Moral Agency of AMAs*; and (iv) *Future Projections about AMAs*.

We acknowledge that, by grouping the statements in these four clusters, we may be failing to include other relevant and interesting topics associated with AMAs. Rather than considering these clusters exhaustive, following the exploratory research tradition, we consider them as baseline ethics disputes surrounding AMAs. Further research may identify and explore other variations and controversies about these artificial systems.

5.2.3. SET OF STATEMENTS (Q-SET)

From the concourse of communication a set of 45 statements was defined (q-set) thus capturing the key disputes and controversies related to AMAs.

Our selection of statements was guided by three main considerations, namely, (i) accounting for a broad scope of positions put forward in the AMAs popular and scientific literature; (ii) favoring clarity; and (iii) avoiding redundancy. Minor edits were made to these statements in order to ensure neutrality and also to meet the number of characters allowed by FlashQ, the software tool that was used in this study for administering the survey.

The q-set reflects the four main clusters mentioned above. More specifically, 14 statements are about the development of AMAs, 18 statements are about the design of AMAs, 8 statements are about moral agency of AMAs, and 5 statements concern future projections about AMAs. Table 5.1 shows a small sample of the statements used in this study. The full q-set (45 statemets) is featured in the Appendix B.

Table 5.1: Sample of statements from q-set (statements 1, 28, 34, and 45).

Development	(1) Technological progress requires artificial morality.
Design	(28) Logic is the ideal choice for encoding machine ethics.
Moral Agency	(34) Because computer programs do not have free will they can never be independent moral agents.
Future Projections	(45) AGI with moral reasoning capabilities will lead to a better understanding of morality.

5.2.4. PARTICIPANTS SET (P-SET)

The target population in this study is (AI) Ethics scholars. The criteria adopted to define this population was having at least one publication in the broad field of AI Ethics. The reasoning behind targeting this population concerns the complex nature of the subject matter, which requires participants to grasp key moral concepts within the context of AI.

Invitations to participate were sent to scholars initially selected through the publications identified in the literature review mentioned above. Subsequently, through snowballing techniques, additional relevant articles and scholars were identified. Each participant was contacted, in the capacity of author or co-author of a particular publication, through the e-mails made publicly available in the publications. Scholars identified in publications in which they were not corresponding authors were also contacted through email, when these were available in personal or institutional websites (invitation e-mail template in Appendix B). This resulted in a large number of invites (n = 277) being sent to (AI) Ethics scholars from June 2020 to December 2020.

A total of 50 participants successfully completed the survey (response rate of approximately 18%). As an inversion of factor analysis that aims just at establishing the existence of particular viewpoints, q-methodology does not require large numbers of participants. In multiple participant q-studies, a p-set consisting of 40-60 participants is considered to be adequate [341, 343].

We believe that the p-set is of adequate size and representative of the target population, which we recall are scholars who have published work in the field of AI Ethics. It is acknowledged, however, that this target population was not rigorously defined. The broad group of (AI) Ethics scholars encompasses several heterogeneous sub-groups (e.g. machine ethicists, robot ethicists, technology ethicists). To disentangle these sub-groups within AI Ethics would be a remarkable effort considering that Ethics scholars often write about various and overlapping topics. For instance, in recent years the Autonomous Vehicle *trolley problem* has been addressed by a myriad of ethicists from different sub-groups of Ethics

[167]. By surveying a heterogeneous group of scholars, we aimed to reach scholars who have written about some but not all the particular topics surveyed, thus ensuring, along with the comments provided by participants about statements ranking highest and lowest, that this study goes beyond and adds value to published literature.

Noteworthy, q-studies do not require a rigorously representative sample but rather a population sample that contains participants with relevant viewpoints on the matter. Whereas selection bias is considered a limitation in confirmatory research, this problem is less salient in q-studies. The objective of these studies is precisely to capture relevant opinions from participants who self-selected to participate in the study. Hence, a selection bias whereby participants with more strong opinions are more likely to participate is not a big problem, as long as sufficient -more neutral- respondents also participate. In that case, all shared perspectives will still be revealed.

We are confident that the p-set in this study includes scholars with relevant viewpoints on the ethical controversies of AMAs. In this context, it should be clarified that a q-study typically makes no claim that the relative sizes of the perspectives (in terms of the number of respondents that adhere to them) reflect the population distribution. A different issue is whether this study succeeded in revealing all perspectives about these controversies. Q-studies should report the shared perspectives about a particular topic. However, it is possible that some scholars who were invited to participate in the study have also strong opinions about AMAs, but failed to complete the survey for technical or personal reasons. This is an unfortunate limitation of this empirical study. Although we believe that the key views about AMAs are duly reported, in future research any extensions and variations from these five baseline perspectives should be investigated.

5.2.5. ANALYSIS

DATA COLLECTION

The data was collected through FlashQ, a software that allows online q-sorting on a grid of columns. The grid was coded as an 11-point distribution ranging from -5 to 5 resembling a simplified bell shaped distribution. The ample range of columns ensures response variability thus allowing participants to reveal nuanced degrees of engagement with different items (Figure 5.1).

Participants were asked to sort the 45 statements according to a subjective notion of disagreement/agreement. The particular arrangements of the items in the grid correspond to the q-sorts of participants. A q-sort represents the perspective of a single participant thus revealing the items that prompt the strongest subjective reactions.

In addition to the (q-sorts) quantitative data, qualitative data was also collected as participants were asked to provide further comments on such statements they ranked in the -5 and 5 columns.

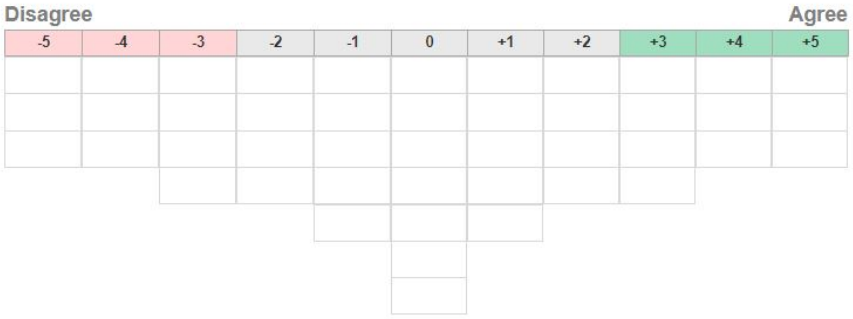


Figure 5.1: Sorting grid.

DATA ANALYSIS

Data analysis in q-studies entails three main steps ((i) factor extraction; (ii) factor rotation; and (iii) factor interpretation). In the analytic process (steps (i) and (ii)) we used PQMethod, a statistical program that accommodates the requirements of q-studies [351]. As a derivation of factor analysis, q-methodology is a data reduction technique which aims to reduce a larger number of variables into fewer factors. Therefore, the analytic process of q-methodology relies on multivariate data-reduction techniques.

The first step in this process entails extracting factors from previously collected q-sorts. Extracting factors consists of summarizing all individual responses into a few representative responses [346]. Either a Centroid Factor Analysis or a Principal Component Analysis can be used for factor extraction. In this study, the factors were extracted through PCA, a linear reduction technique that projects the data into a subspace of lower dimensionality, where the variance of the projected data is maximized, providing the single best mathematical solution [343].

Subsequently, the extracted factors were rotated. Factor rotation aims to position each factor so that its viewpoint closely approximates the viewpoint of a particular group of q-sorts. In PQMethod, this rotation can be done either manually or through an objective solution, which is the Varimax rotation. We used Varimax, an orthogonal rotation of the factor axes that maximizes the variance of each factor loading by making high loadings higher and low loadings lower. Q-sorts that load high on one factor will load low on another, thus maximizing the distinction and differentiation of subject positions while minimizing the correlation among factors [354].

Upon rotating different numbers of factors and comparing the distribution of (automatically flagged) defining sorts among factors, a decision was made to rotate five factors. This solution features the lowest number of factors in which every factor has at least three defining sorts and only one factor has exactly three defining sorts (Table 5.2).

Each factor is characterized by a factor array featuring 45 scores (one score per statement), which is a single q-sort configured to represent the viewpoint of the factor. Given that factors have different numbers of defining sorts, each score in the factor array is a standardized (z) score to allow cross-factor comparison. The five factor arrays are available in Appendix B.

Finally, the interpretation of factors is based on the factor arrays and the comments provided by participants with respect to the statements ranked highest and lowest (comments available in Appendix B). For assisting in the factor interpretation, crib sheets [343] were developed. Crib sheets are useful for displaying the relevant item configuration for each factor thus facilitating the interpretation and analysis of the results. The five crib sheets developed in this study for each perspective feature items ranked -5, items ranked +5, and items that ranked highest or lowest compared to the other array perspectives (crib sheets are available in Appendix B).

Table 5.2: Number of defining sorts in factors [1-5].

	Factor 1	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**
Defining sorts	13	9	8	9	3

5.3. THE CONTROVERSIAL ETHICS OF AMAS

Controversial matters about development, design, moral agency, and future projections for AMAs have been well addressed and debated in the literature [183, 362, 366]. Although a thorough review of the literature is outside the scope of this empirical work, we provide a background on these matters. Building on the main controversies briefly raised in this literature background, we will subsequently present the empirical findings of this study.

5.3.1. ABOUT DEVELOPMENT OF AMAS

As AI systems become increasingly open, decentralized, intelligent, capable of self-regulation, and independent, Machine Ethics looks to run in parallel to ensure that artificial morality is not an afterthought [362, 378].

Machine ethics reasons that the *adjustable autonomy* required to meet AI advances leads to the need for AMAs to deal with, or even replace, human judgment in difficult, surprising, or ambiguous moral situations [362, 379, 380]. It aims to implement ethical principles and moral decision-making faculties in machines to ensure that their behavior towards human users and other machines is ethically acceptable [132, 160]. Further it is claimed that there is value in the pursuit of AMAs, regardless of whether systems are actualized, as it may advance our understanding of morality [127, 128, 381–383].

The Machine Ethics project is, however, quite controversial. The moral admissibility, technical feasibility, and necessity of this project is often questioned [183, 365, 384, 385]. Moreover, whereas Machine Ethics aims to push the boundaries of artificial morality in order to ensure that artificial systems behave ethically, it also opens opens the door to prospects of unethical artificial systems [386] which tend to be detached from state-of-the-art technology and are often dismissed as speculation [365, 387].

5.3.2. ABOUT DESIGN OF AMAS

Designing an AMA is an ambitious scientific and engineering endeavor but it is still unclear whether morality can be reduced to an algorithm [388]. The design of such an advanced artificial moral system primarily entails defining the moral behaviors or ethics that the

system will follow, implementing such moral behaviors or ethics, and operationalizing them.

A preliminary question regarding the moral behaviors or ethics to be implemented in artificial systems, is whether artificial morality should be modeled after human moral competence [380] or if this consideration is altogether a trap [389].

In the scientific literature, several projects attempting to build artificial morality with reference to human morality have been reported. Scholars across different disciplines are exploring the applicability of different branches of moral philosophy, such as duty theories [385, 390–393], consequentialist theories [373, 394, 395], or virtue ethics [363, 384, 396, 397], as well as combinations of several moral theories [146, 398, 399]. Moreover, important work on the modeling of artificial morality based on empirical evidence of human morality, such as the Moral Machine Experiment, has also been reported [133, 134, 271].

The implementation of human inspired moral behaviors in artificial systems typically follows top-down, bottom-up, or hybrid approaches [362, 373, 400–402], but there is no consensus about which implementation is best fit for the endeavor of developing moral machines [403].

Eventually the implemented moral behaviors need to be operationalized so that artificial systems are able to make a determination of a right from a wrong in-situ. In other words, how does a machine moralize? Ethics has to be operationalized so that an AMA is able to recognize a moral situation, weigh up possible moral actions, make moral judgments, and execute them [404].

In the case of designing machines with human-like moral reasoning, there are concerns about the lack of operationalization of the capacities that enables humans to think morally [405]. Although ethics or moral behaviors may be implemented in an AMA through a particular implementation approach, how the AMA operationalizes moral decision-making, and how the designer designs algorithms that account for that process, is still in question and implicates transparency, moral agency, and moral responsibility.

The breadth of different AMA design approaches reported in the literature reveals a lack of consensus among scholars working on the Machine Ethics project and raises questions about whether it is possible to develop an objective validation of AMAs that avoids designer bias and ensures explainability [403, 406–408].

5.3.3. ABOUT MORAL AGENCY OF AMAS

As AI Systems become more autonomous, it has been discussed whether artificial systems ranking high in the autonomy/ethics sensitivity spectrum [362, 375, 404] can be considered to have moral agency. In the indeterministic tradition, moral agency requires personal freedom [409], or at least some sort of faith in personal freedom. Current and foreseeable technology lacks free will, which would therefore preclude machines from having moral agency. However, it is debated whether human-like prerequisites for moral agency should be imposed on machines or if a hard line should be drawn between human moral agency and that of machines [410–414].

Traditionally, a moral agent is an agent that is morally responsible for its actions, i.e. it is attributed certain powers and capacities to be held responsible for exercises of agency

[415]. This description of moral agency is often shared in AMA literature [413, 416]. The debates about attribution of moral agency to AMAs typically entail whether such systems are accountable for their morally relevant decisions.

It has been cautioned in the literature that morally accountable machines may be used to avoid personal responsibility [417]. Hallamaa and Kalliokoski argue that "AI systems are not genuine participants in joint action and cannot be held morally responsible", thus concluding that moral responsibility strictly remains a human characteristic [413].

A *normative turn* with respect to artificial moral agency has also been proposed [371]. That is, to put the discussion about the necessity and requirements for moral agency to the side and move forward to address existing practical needs with AMAs. The debate would therefore be redirected to explore outright normative ethical questions, such as how and to what extent machines should be involved in practices where humans normally assume moral agency and responsibility [371].

5.3.4. ABOUT FUTURE PROJECTIONS FOR AMAS

Looking forward, several domains for AMA implementation are frequently cited, including healthcare, military, and autonomous transport [167, 182, 362, 418, 419]. However, as it often happens with disrupting technologies, the ethics discussions about AMAs tend to be quite speculative.

It is not likely that an artificial agent with such high autonomy and ethics sensitivity as depicted in futuristic narratives is developed prior to Artificial General Intelligence (AGI). So far, however, there is no evidence that that such advanced and generalized forms of intelligence can be developed and it is even questioned if pursuing such research is ethical [372].

There are positive and negative projections about a future with AMAs. They may be our moral teachers or our destroyers [375–377, 420–423]. But for now the societal and moral role of these systems remains unclear.

5.4. PERSPECTIVES ABOUT AMAS

Five main perspectives have emerged from the data collected in this empirical study, thus illustrating the heterogeneity of opinions about AMAs. These perspectives are described with reference to the four themes associated with AMAs which were identified above. Each perspective features an array of 45 scores (Appendix B) in which the score assigned to each statement results from a standardization process of the scores that the participants who loaded significantly on that perspective assigned to that particular statement. The core characteristics of each perspective are derived from the statements ranked -5 and +5 [$(N :\mid 5 \mid)$ where N is the number of the statement and $(\mid 5 \mid)$ may be either -5 or +5] as well as the statements ranked highest or lowest compared to the arrays of the other perspectives [$(N :\mid P_i \mid)$ where N is the number of the statement, P_i is the perspective with $i \in [1, 5]$, and $\mid P_i \mid$ may either be $-P_i$ or $+P_i$ depending if the statement is ranked lowest or highest than in the arrays of other perspectives]. These perspectives are summarized in Table 5.3 and further characterized and discussed in subsequent sections.

Table 5.3: Five perspectives about AMAs.

→ Themes ↓ Perspectives	Development	Design	Moral Agency	Projections
P1. Machine Etics: The way forward	Unavoidable & permissible	Moral certainty Ought ← is	Potential for agency	Positive
P2.Ethical Verification: Safe & Sufficient	Not required	Verification & Governance	Human agency ≠ AI agency	AMAs will not be our moral teachers
P3. Human values to avoid moral dystopia	Unavoidable & permissible	Moral uncertainty	Potential for agency	Possible existential threat
P4. Human Exceptionalism: Machines cannot moralize	Ethics not reducible to algorithms	Logic not a good choice for encoding morality	AI cannot achieve moral agency & Free will is required	Skepticism & AMAs will not be our moral teachers
P5. Machine Objectivism: Machines as superior moral agents	Needed to prevent harm	Logic	Potential for agency & Free will is not required	Machines will be better moral agents

5.4.1. PERSPECTIVE 1 *Machine Ethics: The Way Forward*

AMAs are unavoidable and may be necessary for technological progress. Moral sureness and a mixed design approach is key, but free will is not. More than simple tools, AMAs will advance our understanding of morality.

Development of AMAs According to this perspective, autonomous moral machines are unavoidable (2:+5) and they might even be a requirement for technological progress (1:+P1). Two participants who loaded significantly on this perspective elaborated further on these topics. One participant wrote that *Technology we already have such as search algorithms and driverless cars require and implement value judgments to a limited extent, primarily as directed by human input, but there are already hints that these limitations can and will be surpassed to at least some extent.* And another participant remarked that *There will be no other way than to develop ethical machines when humanity is supposed to rely with their life on them.* Moreover, as per this perspective, creating AMAs is permissible according to existing moral theories (11:-5) and will increase public trust and confidence in creating autonomous agents acting on behalf of humans (8:+P1).

Design of AMAs Moral sureness is valued when it comes to AMA decisions (18:-5), but a mixed approach (top-down, bottom-up, supra-rational) is accepted in arriving at those decisions (30:+P1) (*This seems like the most viable path forward for AMAs as it allows for context specificity, adaptive response, and learning*). As per this perspective, deriving an ought from an is, by implementing social preferences in machines, is not seen as problematic in AMA design (17:-P1) (*To pretend that we can create machines that lack our biases and are uninfluenced by our values is misguided*).

Moral Agency of AMAs AMAs cannot be understood as simple tools, given the potential for agency (37:+P1) (*Machines are already at least at the level of animals, which are capable of not merely being restrained but being trained*). Since humans may also lack free will, according to this perspective, free will is not essential for moral agency (34:-5) (*It is very likely that even humans lack radical free will. If we require moral agents to have free will, then there will be no moral agents*).

Future Projections A positive outlook about the Machine Ethics project is observed in this perspective, as it holds that developing AMAs and ultimately AGI with moral reasoning capabilities will lead to a better understanding of morality (9 & 45:+5) (*I believe that implementing process-level models of such theories and testing them in various situations is an invaluable method for evaluating said theories*).

5.4.2. PERSPECTIVE 2 *Ethical Verification: Safe and Sufficient*

AMAs will not replace humans in ambiguous moral situations as ethics and human moral agency are not algorithmic. Transparency, accountability, and predictability leads to sufficiently ethical machines.

Development of AMAs In the second perspective identified in this study, technological progress will not require artificial morality (1:-P2) (*Technological progress can and should be guided by ethical and societal considerations and can happen also without artificial morality*).

Moreover, AMAs are not expected to replace humans in difficult, surprising, or ambiguous moral situations any time soon (42:+5). Three participants provided relevant comments about this statement. (i) *At present AMAs have only been demonstrated in laboratory tests, of limited scope ... There is a huge gap between the capabilities of present day minimal AMAs and the ability of humans to make judgements in ambiguous moral situations ... Closing that gap will take many decades of research, and might even prove impossible without fundamental breakthroughs in AGI and machine consciousness.* (ii) *Even if machines are capable of making moral decisions, completely replacing humans in such situations might lead to responsibility gaps.* (iii) *I now believe we are far from having morally competent agents, and that the threats from lack of transparency, privacy, security, etc. are far more pressing, morally speaking.*

One practical limitation to artificial morality, according to this perspective, is that ethics cannot be reduced to a computational algorithm (14:+5) (*Moral judgements, sentiments and motivations depend on a holistic perception of the world. Ethics would not exist at all without this special perspective that is shaped by reasons, emotions, culture and history. A representation of ethics in the form of computational algorithms (or in any kind of model) is an abstraction in comparison to the rich features of the ethical world. Such a representation may successfully serve a specific purpose when realized in a technological artifact, but no representation could possibly model the whole ethical world as a subset of its features*).

Design of AMAs Rather than expecting AMAs to grasp moral principles (21:-P2), they should be moderated through the verification of transparency, accountability, and predictability (32:+5) (*If an AMA makes the wrong decision the outcomes could be disastrous. Similarly the risks of malicious hacking of AMAs are serious. This verification, validation and transparency are critical to the success of (even limited) AMAs in real world use. Equally important is very strong governance of AMAs, so that their ethical performance is carefully monitored, and both accidents and near-misses thoroughly investigated*).

Moral Agency of AMAs There exists an essential difference between human and artificial moral agency (39:-5), namely phenomenal consciousness and currently unknown cognitive processes relating to human reality. These features, among others which constitute moral agency, are not quantifiable (40:-5) ((i) *There is a vast difference. We do not understand the cognitive processes of human morality - which likely depend on both rational and emotional responses, alongside experience. In contrast AMAs are based on simple abstract models, which are far from even scratching the surface of human ethical judgement.* (ii) *Ultimately consciousness is of concern here, and specifically phenomenal consciousness, since the functional parts of consciousness are becoming better understood ... There is no first-person perspective for artificial systems replete with experiential properties*).

It is not inevitable that machines will become full ethical agents (3:-5) ((i) *It seems possible that there might either be moral or political grounds for stopping the development of AMAs or just technological inability.* (ii) *I would agree that (many) machines would inevitably have ethical impact, but I don't believe that they should be full ethical agents, with the implication that this would mean replacing humans.* (iii) *A full ethical agent would be one that perceives the world in a holistic way, shaped by reasons, emotions, culture and history etc. It would have to grow up and 'live' in a process of constant evolvement in relationships with people, with society, with culture, history, religion etc. Although this might not be considered*

impossible if taken up in a though experiment, it makes no sense in reality).

Future Projections According to this perspective, AMAs, even if endowed with human-centred values, will not play a role in educating humans on morality (43:-P2).

5.4.3. PERSPECTIVE 3 *Morally Uncertain Machines: Human Values to Avoid Moral Dystopia*

AMAs must be morally uncertain and hold human values otherwise they pose an existential threat. Simply prohibiting unethical behavior, as well as implementing external checks and balances, is not enough.

Development of AMAs As per this perspective, AI Systems in morally salient contexts will not and cannot be avoided (5:+5) (*We already see examples with self-driving cars and trading bots, but more generally I see it as (nearly?) inevitable that AI systems will eventually be deployed in every domain that requires intelligence, which is essentially a superset of all domains that contain morally salient situations*) and the creation of moral machines is morally permissible according to existing moral theories (11:-5).

Design of AMAs On the design of intelligent machines, it is not enough to restrict ethical concerns to the prohibition of unethical behavior (15:+5) (*Negative ethical restraints will not be sufficient. Many social and ethical issues and progress itself require careful deliberation and proactivity*).

It follows that external checks and balances such as safety features, codes of conduct, certification processes, and clear limits to the operational contexts are not sufficient to ensure machines will not hurt humans (10:-5).

Unlike P1, which did not value moral uncertainty and instead favored moral sureness in AMA decisions, this perspective values machines being fundamentally uncertain about morality (18:+P3) (*Morality is a critical determinant of ethical behavior, and there is incredible disagreement among humans. If AGI does not have uncertainty about morality, its behavior may be arbitrarily bad given a commitment to the wrong set of moral principles*).

Moral Agency of AMAs This perspective rejects the idea that artificial moral agency will remain a primitive form of moral agency compared to that of human beings (35:-P3) ((i) *It is currently primitive, but I believe it will eventually be possible to create AIs that match or exceed humans in every intellectual capability, which includes moral reasoning/agency. (ii) Artificial moral agency can become the paradigm of ethics, and is not necessarily bound to remain a lesser, more mechanical form of assessing and relating to situations*).

Free will is not required for moral agency and machines lacking free will can be independent moral agents (34:-5) ((i) *there is no free will in the libertarian sense ... Humans have designs, like machines do, so the fact that machines and do or do not have free will and that they are designed is not especially salient to the question of moral agency. (ii) ... if human beings can be said to have free will under any particular definition, then it is possible to implement a program that can be said to have free will under that particular definition*).

Future Projections Mere indifference to human values—including human survival—could be sufficient for AGIs to pose an existential threat (44:+5) (*Thought experiments such as the paperclip maximizer show quite convincingly that for an AGI to pursue a goal that is merely orthogonal to human values could plausibly present an existential threat*).

5.4.4. PERSPECTIVE 4 *Human Exceptionalism: Machines Cannot Moralize*

AMAs are without moral agency as they lack free will and the understanding to make moral assessments as humans do. Logical machines will not be better moralizers or our moral teachers.

Development of AMA According to this perspective, ethics cannot be reduced to a computational algorithm (14:+5) ((i) *Ethics is not about calculations, but about not quantifiable preferences.* (ii) *Mostly because ethics, or at least what ethics deals with, requires a plurality of points of view related to the physical embodiment and location of independent agents, which means that there is no possible universal description of an embodied agent's position or situation. Even considering an embodied AMA the difficulty is that an agent only acts morally if he or she or it could have acted immorally. We could simulate that by including a random generator but acting immorally is not the same thing as acting randomly*).

Unlike the other perspectives that held strong positive views about the permissibility of creating moral machines according to the tenets of existing moral theories, P4 is quite neutral about it (11:+P4) (*If it turns out that it is best to have moral machines, then utilitarianism would permit and even require that we bring about moral machines*).

Design of AMAs With respect to the design of AMAs, logic is not considered the ideal choice for encoding machine ethics (28:-P4).

Moral Agency of AMAs Humans and AMAs are not alike as far as moral agency is concerned (39:-5) (*Without sentience computers cannot express agency*) and machines will not inevitably become full ethical agents (3:-5) ((i) *Agency requires consciousness. (ii) The machines we can now produce certainly are not and the planned AMA I know of will certainly not be full or real moral agents though there is no reason to think that it is in principle impossible*).

Only this perspective indicated that computers lack the conceptual understanding to make a moral assessment which precludes them from achieving moral agency (33:+5) ((i) *These are machines. Very complex, but machines. Yes, some programmer can shape these machines to have an input/output function that produces a behaviour that some human observer may see as analogous to human moral behaviour, but computing machines are ultimately and intrinsically incapable of understanding in human terms, which is the basis for moral agency.* (ii) *You start being moral when you recognize your shared humanness with others and understand that, like it or not, you are in relationship with them. Until machines get that (and I'm suspicious of their ability to do so) then they're not going to have full moral agency.* (iii) *What is it like to be a computer? If there is nothing that it is like to be a computer then how can a computer have conceptual understanding?* (iv) *Computer lacks empathetic experiences which give humans the conceptual understanding needed for moral agency*).

Also uniquely positive in this perspective was the agreement that computer programs can never be independent moral agents as they lack free will (34:+P4) (*Free will strikes me as a basic condition of responsibility and therefore of moral agency*).

Future Projections There is a long way to go before artificial agents can replace human judgment in difficult, surprising, or ambiguous moral situations (42:+5) (*Unexpected situations are hardly manageable artificially*) and AMAs will not be our moral teachers (43:-5) ((i) *[Morality] is about convictions and convergence or clash. Nothing to teach there, at least not from a machine.* (ii) *Unless machines develop empathy and compassion, they aren't really going to pass down lessons in a meaningful way.* (iii) *Morality cannot be taught simply through information transfer: it requires experiential sharing*). Moreover, AGI with moral reasoning capabilities will not lead us to a better understanding of morality (45:-P4).

5.4.5. PERSPECTIVE 5 *Machine Objectivism: Machines as Superior Moral Agents*

AMAs prevent human harm. Through logic and context-specificity, they are better moral reasoners and educators. Free will and conceptual understanding are not required for moral agency.

Development of AMAs In this perspective, unlike all others, there is a strong view that the development of AMAs prevents machines from hurting human beings (7:+P5).

Design of AMAs This perspective challenges the notion that machines should use societal preferences to identify an informed and desirable choice when faced with a specific ethical dilemma (27:-5) (*I believe that machines can enhance us as moral agents if they manage to distance us reflectively from our intuitions, which are very much determined by social preferences*). Moral implementation strategies should be context-specific (13:+5) and logic rather than common sense (20:-P5) is the best choice for encoding machine ethics (28:+P5).

Moral Agency of AMAs Conceptual understanding and free will are not considered necessary conditions for moral agency and so moral agency, even if primitive, may be ascribed to machines (33 & 34:-5).

Future Projections On future projections, developing AI Systems and AGI with moral reasoning capabilities will ultimately lead to a better understanding of morality (9 & 45:+5). It is projected that machines will be better moral agents than humans since they are not subject to irrationality, seduction, or emotional turmoil (41:+P5).

5.5. DISCUSSION

5.5.1. CONTRASTING VIEWS & AGREEMENTS

The five different perspectives about AMAs identified in this empirical study reveal contrasting views on artificial morality. Particularly salient differences between perspectives arise with respect to the development and moral agency of AMAs. About the development of AMAs, *Machine Ethics: The Way Forward* (P1), which stands for advancing artificial

morality, is in sharp contrast with *Ethical Verification: Safe and Sufficient* (P2), which is skeptical about the feasibility or need for artificial morality. As for moral agency of AMAs, *Human exceptionalism: Machines Cannot Moralize* (P4), which values the human aspect in morality and therefore does not accept that computer programs can be moral agents since they lack humanness, contrasts with *Machine Objectivism: Machines as Superior Moral Agents* (P5), which views that morality improves when stripped of human flaws. In addition to the differences between perspectives, there are also transverse contrasting views and agreements to be reported with respect to the key matters explored in this study.

Regarding the development of AMAs, most perspectives agree that AI systems working in morally salient contexts cannot be avoided and, as such, some degree of moral competence ought to be demonstrated. However, differences arise as to whether that is the moral competence of the machine or the designer. There is also a general consensus about the permissibility of developing AMAs as per existing moral theories. Whether AMAs ought to be developed, even if feasible and permissible according to moral theories, is a different point of contention.

On the contrary, the design of AMAs is a fracturing topic with different perspectives favoring different approaches. Design approaches based on societal preferences, which derive an ought from an is, as seen in the Moral Machine Experiment [134], also divided the perspectives. It is no surprise that discussions on how to design for ethics yield disagreement. As an ethic, machine ethics is susceptible to different opinions, perspectives, experiences, and worldviews of contributors in the field. As an applied ethic, which concerns the application of normative ethical theories to problems with practical limitations and expectations, Machine Ethics is at the mercy of not only the philosopher but also those working in the fields affected and the state of the field itself.

Ideas on moral agency are also diverse. There is no agreement about the moral agency status of AMAs today or in the future. However, with the exception of P4, some consensus is reported about free will not being essential for moral agency. This position emerges from the thought, shared by participants in this study, that provided that humans do not have (at least radical) free will and yet are moral agents, the same should apply to machines. Provided that a not insignificant number of philosophers commit to libertarian views of free will, we speculate that this consensus may reflect more the background of participants rather than traditional philosophical ethics. Future empirical research studies engaging a larger set of ethicists should further explore this issue.

About future projections, there is an overarching agreement that there is a long way to go before AMAs replace human beings in difficult moral situations. However, the future societal and moral role of these highly advanced artificial systems is mixed. Our data shows uncertainty about whether such systems will be superior moral reasoners, avoid a moral dystopia, or lead to a better understanding of morality.

In order to contextualize our findings in the broad scope of AI and Machine Ethics, we further reflect on the source and implications of the differing views and baseline agreements on the key matters about AMAs reported in this study.

5.5.2. CONTRASTING VIEWS & THE FAILURE OF MACHINE ETHICS AS AN APPLIED ETHIC

A potential source of these differing perspectives is the failure of Machine Ethics to be widely observed or explored as an applied ethic and more than a futuristic end. AMAs exist only in laboratories and are mostly intangible at present. The Machine Ethics literature is therefore chiefly abstract and fails to move beyond normative and descriptive ethical theory with the examination of consequentialist machines, social norm machines, virtuous machines, etc.

As remarked in some of the perspectives identified here, AMAs are often presented as outsiders, as superior moral reasoners following programmed, taught, or learned normative or descriptive ethical theory. Moreover, perceptions about future projections for AMAs are unsurprisingly mixed since AMAs are by and large objectively unattainable, given that current technology and future advancements are unknowable.

Consequently, the practical discussions about the ethics of particular AI Systems are realized primarily in relation to the field affected, e.g. AI in healthcare or AI in transportation [167, 317, 318]. In these discussions, the ethics concerning the AI system is applied and shaped to the field in which it will operate and not the other way around.

It is therefore speculated that AMAs are yet to be widely accepted in the same way AI systems have been because they have emerged in the literature as not only artifacts of the future but as outside and superior enablers, advisers, learners, or lecturers of ethical theory without any or much regard for the field. For Machine Ethics to be an applied ethic, AMAs ought to be shaped to the present-day expectations, norms, codes, and stakeholders of the field in which it intervenes.

From the failings of Machine Ethics to be widely regarded as an applied ethic and a feasible pursuit, a second main perspective with respect to development and future projections of AMAs (*Ethical verification: Safe and Sufficient*) emerges in response. It reflects the views of practitioners who often see the Machine Ethics project as unattainable, futuristic, and disconnected from the practical domain. Developing checks and balances at either the higher level, like policy, or within machines, such as implicitly safe or human operator takeover mechanisms is presented as an adequate solution for keeping autonomous machines in check.

Looking forward, AMAs might leave the laboratories of a few select Machine Ethics researchers to be widely developed and accepted, but it is just as likely that we will continue down the existing path of building safe systems designed with an enormous amount of ethical consideration and interdisciplinary input. Yet, in the meantime, there may be something to be learned from the pursuit of AMAs with an interdisciplinary approach. Machine Ethics research could lead to new insights into human morality and, at the same time, it can be grounded by practitioners who can help to guide the realization of moral machines in the field.

5.5.3. AN INFORMED DEBATE ON AMAS

The starting point to realize an interdisciplinary approach to AMAs is an informed debate where contrasting views and agreements are disclosed and appreciated. By systematically reporting different perspectives about AMAs, we believe our exploratory research lays the

foundations for such debate.

Further clarity about positions held in AMA contributions on the disputes surrounding AMAs could be realized if researchers make explicit their views about the development, design, moral agency, and future projections for AMAs. This could be done at the start of works to prime the reader, making the views and interpretations held in the contribution plain and enabling a well-informed reading of the material.

An informed debate also facilitates the identification of theoretical and practical research opportunities. Our study indicates that further research is needed to outline the relation between free will and moral agency of artificial systems, which could lead into new insights about moral agency. By clearly reporting the contrasting views with respect to the design of artificial morality, we also identify an opportunity for practitioners to weigh in on ethical design and propose their own (grounded) solutions. Moreover, we expect that the marginal agreements reported in this research, about the inevitability of AI in morally salient contexts and the need for moral competence, are further explored and developed into shared research principles.

5.6. CONCLUSION

This empirical study explored the controversial topic of AMAs and aimed to establish an informed debate where contrasting views and agreements are disclosed and appreciated. For this purpose, fifty (AI) Ethics scholars were surveyed. The results empirically demonstrate the wide breadth of viewpoints and approaches to artificial morality.

Although an effort was made to capture the disputes and controversies surrounding AMAs in the popular and scientific literature, it is acknowledged that the four central themes in this research (development, design, moral agency, future projections) and corresponding statements fail to account for every dispute or controversial thought about AMAs. In the exploratory research tradition, rather than considering these four themes and clusters of statements exhaustive, we consider them as baseline ethics disputes surrounding AMAs.

Five main perspectives about AMAs have emerged from our data, thus providing further insight about the disputes surrounding the development, design, moral agency, and future projections for these systems ((i) *Machine Ethics: The Way Forward*; (ii) *Ethical Verification: Safe & Sufficient*; (iii) *Morally Uncertain Machines: Human values to Avoid Moral Dystopia*; (iv) *Human Exceptionalism: Machines Cannot Moralise*; (v) *Machine Objectivism: Machines as Superior Moral Agents*).

The diverse perspectives identified in this study have implications for the Machine Ethics project, which is primarily perceived as either the best way forward to realize ethical machines or as futuristic and lacking practical application of moral considerations. Upon analysis of the perspectives that emerged in the data collected in this empirical study, it is hypothesized that a potential source of these differing perspectives is the failure of Machine Ethics to be widely observed or explored as an applied ethic and a feasible pursuit.

In order to realize an interdisciplinary approach to artificial morality, which allows us to gain insights into morality while also engaging practitioners, an informed debate about AMAs is crucial. Our study helps improve the foundations for such debate. It opens

avenues for further clarity about views on the development, design, moral agency, and future projections in AMAs research and facilitates the identification of theoretical and practical research opportunities.

6

COMPUTER SAYS I DON'T KNOW: AN EMPIRICAL APPROACH TO CAPTURE MORAL UNCERTAINTY IN ARTIFICIAL INTELLIGENCE

A. Martinho, M. Kroesen, and C. Chorus, *Computer Says I Don't Know: An Empirical Approach to Capture Moral Uncertainty in Artificial Intelligence*, Minds and Machines **31**, 215 (2021).

ABSTRACT

As AI Systems become increasingly autonomous they are expected to engage in complex moral decision-making processes. For the purpose of guidance of such processes theoretical and empirical solutions have been sought. In this research we integrate both theoretical and empirical lines of thought to address the matters of moral reasoning in AI Systems. We reconceptualize the metanormative framework for decision-making under moral uncertainty as construed by William MacAskill and Kyle Bogosian within the Discrete Choice Analysis domain and we operationalize it through a latent class choice model. The discrete choice analysis-based formulation of the metanormative framework is theory-rooted and practical as it captures moral uncertainty through a small set of latent classes. To illustrate our approach we conceptualize a society in which AI Systems are in charge of making policy choices. In the proof of concept two AI systems make policy choices on behalf of a society but while one of the systems uses a baseline moral certain model the other uses a moral uncertain model. It was observed that there are cases in which the AI Systems disagree about the policy to be chosen which we believe is an indication about the relevance of moral uncertainty.

6.1. INTRODUCTION

The inner workings and innuendos of morality remain obscure yet the design of a moral compass for Artificial Intelligence (AI) Systems is pressing. The modern AI System benefits from robust computational power and sophisticated algorithms to feed on data for its own learning and adaptive processes thus becoming increasingly autonomous while, at the same time, engaging in complex decision-making processes. Such complexity is aggravated by the potential moral dimension of decisions and there are concerns about whether these systems will uphold moral values [424].

A number of cases where decisions made by AI Systems have morally problematic implications have been discussed in the literature. The archetypal case is the autonomous vehicle (AV) moral dilemma, a philosophical situation, modeled after the *trolley problem* thought experiment [251, 252], in which the AV is required to make a moral choice between actions in traffic that will result in different combinations of lives saved and sacrificed [134, 140, 142, 163, 232–234]. In the services industry it was reported that a machine learning algorithm used by Amazon, which was eventually dropped, penalized female applicants [100, 410]. Another well explored case is related to the use of COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a machine learning based algorithm used in recidivism risk calculations for parole and bail in the U.S justice system, which was reported to be racially biased [86, 103, 410, 425, 426].

Theoretical and empirical lines of thought have emerged in the literature to resolve issues associated with moral reasoning in and by AI systems. These researches shape Machine Ethics, a controversial ethical field, which aims at implementing ethical principles and moral decision-making faculties in machines to ensure that their behavior toward human users and other machines is ethically acceptable [132, 160, 312, 365, 366].

In theoretical work, there appears to be a general consensus that conceptual agreement regarding the moral machinery of artificial agents should precede design endeavors. This has led to a renewed interest in normative ethics work that spans several domains of knowledge and research. Prospects were made for moral machines based on moral theories, such as deontology and consequentialism [127, 160, 182, 373, 385, 390, 427] but the problem of moral disagreement between competing moral theories and conflicting moral judgments was never surmounted [398]. A solution that has been advanced in more recent literature is to design AI Systems to be fundamentally uncertain about morality [398, 401, 428].

Decisions made by these systems within the realm of moral uncertainty would be based on the assumption that there is no certainty about which moral theory is correct. A particular theoretical framework for decision-making under moral uncertainty developed by William MacAskill has been outlined within the domain of AI morality by Kyle Bogosian [398, 429]. It is based on the metanormative notion that moral uncertainty can be conceived as a voting problem among moral theories. The resulting moral preference of the AI is then a function of the credence in particular moral theories weighted by the moral acceptability ("choice-worthiness") of an action under each theory. This translates into an ordering of actions that maximizes overall choice-worthiness [398, 429–431]. Because this framework treats moral uncertainty as a voting problem among moral theories it overcomes common objections of incomparability and incommensurability of moral theories and moral judgments. By allowing AI systems to act in accordance with the diverse values endorsed by humans, the system accommodates the diversity of moral values in its moral decisions [398, 429, 430,

432, 433].

Empirical attempts to address and embed moral reasoning in AI Systems rely on the idea that human morality should be reflected in AI moral reasoning. Human morality would first need to be empirically identified and subsequently embedded in the AI system [134, 144, 147, 163, 434, 435]. In the Moral Machine Experiment, a particularly impressive empirical research endeavor in which preference data of 1.3 million respondents from various regions of the world was compiled in the context of moral dilemmas, it was suggested that the relative agreement found is a positive indicator for consensual morality in AI [134]. Although this notion of moral agreement is indeed attractive from a pragmatic viewpoint, it has also raised some criticism [150, 436]. Other studies have made attempts to capture moral heterogeneity across individuals, inspired by the above mentioned line of theoretical argumentation. This has, however, proven to pose severe practical challenges in terms of empirical operationalization. Most pressingly, capturing every preference or vote gives rise to practical runtime problems (which could be particularly problematic in cases where the AI has to make split-second decisions), yet, averaging preferences or votes into one preference profile [271] comes with the risk of failing to properly account for marginal preferences.

This paper contributes to the theoretical and empirical strands of literature, which focus on embedding moral values and judgments into AI systems, by providing a theory-rooted yet empirically practical approach to capture society's moral heterogeneity in a morally uncertain AI system. Our approach builds on the current theoretical understanding that moral uncertainty is paramount to the endeavor of implementing moral reasoning in AI Systems and it is practical by avoiding runtime issues and not requiring averaging efforts. We propose to generate such moral uncertainty by re-conceptualizing and operationalizing the metanormative framework for decision-making under moral uncertainty, briefly introduced earlier, as a utility-based latent class discrete choice model.

Moral heterogeneity is captured through a small set of latent classes, each with its own distinct moral preferences, which makes this theory-rooted approach for moral decision-making of AI systems practical in terms of runtime and interpretability. Without loss of generality we use a small-scale dataset that resulted from a choice experiment to provide an illustration in which an AI System makes policy choices on behalf of societies based on the conflicting moral preferences of latent classes in society.

The novelty of this work is the use of discrete choice analysis to codify human (moral) preferences and decision rules in order to embed these into a (moral) AI System and, moreover, an empirical illustration of moral uncertainty. With this research we expect to contribute to the Machine Ethics and Artificial Moral Agents (AMA) literature [127, 132, 160, 162, 183, 366, 437], as well as the moral uncertainty literature, which has mainly explored theoretical case studies [431, 438], and also to broader lines of research emphasizing the need to embed values into AI [439, 440].

The remainder of the paper is organized as follows. Section 6.2 explains at a conceptual level how the recently proposed metanormative framework can be connected to Discrete Choice Analysis. Section 6.3 goes further by showing how an operational latent class discrete choice model can be used to codify moral uncertainty in AI. Sections 6.4 and 6.5 illustrate the approach in the context of a concrete example, where the latter section focuses on estimating the latent class model on choice data. Section 6.6 presents the proof of concept,

by equipping AI systems with a moral framework under the assumptions of normative certainty versus normative uncertainty, building on the modeling efforts presented in preceding sections. Section 6.7 draws conclusions and puts forward avenues for further research.

6.2. A METANORMATIVE FRAMEWORK AND ITS CONNECTION WITH DISCRETE CHOICE ANALYSIS

The metanormative framework, as construed by MacAskill and Bogosian, produces an ordering of actions in terms of their choice-worthiness in a particular decision-situation [398, 429]. The key elements in the decision-situation, i.e. a context in which an agent is required to make a decision, are the *decision-maker*, which in this research is an AI System defined as a regularly interacting or interdependent group of units which form an integrated structure that employs AI in any of its forms (learning, planning, reasoning, natural language processing, perception) separately or combined to perform a function while continuously interacting with the environment [441]; a set of possible *actions* that the decision-maker has the power to bring about; the *normative theories* taken into account by the decision-maker; a *credence function* that represents the decision-maker's beliefs or trust in the various normative theories; and the *choice-worthiness*, which is the normative ordering of actions after all relevant considerations have been taken into account [398, 429].

At the core of this metanormative framework for capturing moral uncertainty is the notion that the choice-worthiness of an action is determined by its choice-worthiness according to various competing normative or moral theories and the credence of the decision-maker in each of those theories. More precisely, the choice-worthiness of an action is the credence-weighted average of the choice-worthiness of the action in all of the individual theories. Using a slightly adapted notation, this core can be formalized as $W(a_i) = \sum_t^T [C(t) \cdot W_t(a_i)]$ where $W(a_i)$ denotes the total or over-all choice-worthiness of an action; a_i is an action from the exhaustive set A of mutually exclusive actions $\{a_1 ... a_i ... a_J\}$ where J is the cardinality of the choice set; $W_t(a_i)$ denotes the choice-worthiness of an action given a particular normative theory t which is taken from the set T of available theories; and $C(t)$ denotes the credence of the theory[1].

The operationalization of this formulation entails two important challenges regarding the measure or inference of $W_t(a_i)$, i.e. the choice-worthiness of an action given a moral theory, and the measure or inference of $C(t)$, i.e. the credence of a moral theory. We present Discrete Choice Analysis as an intuitive method to make these inferences in an empirically rigorous way. A reconceptualization of the formulation introduced above is required so it can be re-positioned into the Discrete Choice Analysis domain. Firstly the choice-worthiness of an action given a moral theory is re-conceptualized as the utility of an action given a moral theory. Although this variation is in fact a matter of semantics, it facilitates the connection with the micro-econometric framework of Discrete Choice Analysis. Further details on the definition and operationalization of the concept of utility

[1] It is acknowledged that this formulation is a restricted version of the metanormative framework proposed by MacAskill and Bogosian in that their original framework features various important extensions to account for different types of moral theories. For now, however, the focus of this research is on what it is believed to be the core of the metanormative framework thus opening an avenue for further research that accommodates its various extensions.

will be provided in the upcoming sections. A second and more relevant step is the re-conceptualization of the credence of a theory into the share of the population that adheres to that theory or, equivalently, the probability that a randomly sampled individual from the population adheres to the theory. It is therefore implicitly postulated that $C(t) \in [0,1] \forall t$, and that $\sum_t^T C(t) = 1$ which is in fact congruent to the construction of credence in the metanormative framework as an "assignment of probabilities to various moral theories of being correct"[398].

To avoid confusion, a new notation (V for utility of an action and P for the probability that a sampled individual adheres to a theory) is adopted thus leading to the following formulation for the utility of an action: $V(a_i) = \sum_t^T [P(t) \cdot V_t(a_i)]$. The challenge is to measure or infer $P(t)$ and $V_t(a_i)$. As will be elaborated below, the domain of Discrete Choice Analysis, and its sub-branch of latent class discrete choice modeling, offers a powerful approach to tackle this challenge. As an empirical base, we use experimental choice data (observed choices made by human decision-makers in a choice situation) to estimate the probability that an individual belongs to a specific class associated with a particular moral theory $P(t)$ as well as the utility of an action a_i given a moral theory $V_t(a_i)$.

6.3. OPERATIONALIZATION OF THE METANORMATIVE FRAMEWORK USING DISCRETE CHOICE ANALYSIS

As mentioned above, the model that will be used to operationalize the re-conceptualized metanormative framework is drawn from Discrete Choice Analysis. The key idea in this field is that a choice provides a signal of the latent utility of the choice options or alternatives [442, 443] and that the utility that the decision-maker ascribes to each alternative in the choice set can be inferred by means of econometric methodology [444–446]. For this purpose, a choice model has to be built that explicitly relates utilities to choices in such a way that utilities can be inferred from the choices in a statistically rigorous process.

In a broad sense, discrete choice theory generates mathematical models that formally describe and explain the decision-process of an agent or a group of agents that make a choice between two or more mutually exclusive discrete alternatives from a finite choice set. A choice model is defined in terms of the input variables it includes, their associated parameters, a decision rule, and an error structure. These models are probabilistic in the sense that they generate choice probabilities for each alternative in a choice set. This means that choice models reflect not only that decision-makers are to some extent inconsistent or random in their behavior but also that the model does not capture all information that may be relevant for every single choice, as well as the fact that preferences differ across individuals.

We present a brief notation of a choice model to elucidate the relation between latent class based discrete choice theory and the re-conceptualized metanormative framework. The model will be based on Random Utility Theory, which assumes that the utility of an action a_i is the sum of observable (deterministic or systematic utility) and unobservable (random utility) components of the total utilities: $U_{in} = V_{in} + \epsilon_{in}$ [447–449].

The unobservable component in Random Utility Theory is an error term that captures noise. Although we will not elaborate on the intricacies of this disturbance term, in the remainder of this paper it is important to note that depending on the assumed distribution

of the term, there are different formulations of the probability that a randomly sampled individual from the population chooses a_i from the set of actions A. The by far most used formulation is the so-called Logit function which assumes that errors are extreme value distributed type I with variance $\frac{\pi^2}{6}$. In this formalization of the error term, the probability that a_i is chosen is written as: $P(a_i) = \frac{exp(V(a_i))}{\sum_{j=1}^{J} exp(V(a_j))}$. In a process of maximum-likelihood based optimization the utilities are obtained for each alternative which together (through their implied choice probabilities) make the observed choices the most likely. The result of this process is an econometric estimate $\hat{V}(a_j)$ for every alternative j in A, including i.

Whereas conventional choice models, like the Logit model introduced above, implicitly assume that the utilities of the population can be represented in one single estimate for each alternative, the latent class approach alleviates this restrictive assumption by postulating that there may exist several latent classes in the population, with homogeneous utilities within each class, which are different from those in other classes. In other words, the latent class choice model is based on the assumption that a number of segments or classes exist within the population featuring different preferences albeit internally relatively homogeneous [450]. These models provide insights into heterogeneity of preferences and decision rules of people while accounting for the fact that different segments of the population have different needs and values and, in consequence, may exhibit different choice preferences. Since it is not known *a priori* which decision-makers belong to each class, the segments are treated as latent rather than predetermined by the analyst. This means that the choice model provides a solution to the problem of unobserved heterogeneity. It determines simultaneously the number of (latent) segments and the size of each segment and it also estimates a separate set of utility parameters for each segment [451–454].

The choice probability given by a latent class choice model is written as: $P(a_i) = \sum_t^T [P(t) \cdot P_t(a_i)]$. This means that the probability that a_i is chosen is the probability that a randomly sampled individual belongs to a class t (this is called the class membership probability and is denoted by $P(t)$ for individuals in class $t \in T$ where T is the set of all mutually exclusive and commonly exhaustive classes of decision-makers[2]) multiplied by the probability that a_i is chosen by a decision-maker from a particular class t ($P_t(a_i)$), summed over all classes. Estimation of such a latent class choice model results in not only an estimate of the probability that a randomly sampled individual belongs to a class (i.e., the share of the population that belongs to that class) ($\hat{P}(t)$), but also an estimate of class-specific utility for each alternative j and for each class t ($\hat{V}_t(a_j)$).

Revisiting re-conceptualized metanormative formulation introduced above $V(a_i) = \sum_t^T [P(t) \cdot V_t(a_i)]$ it is now clear that through Discrete Choice Analysis econometric estimations of the two crucial elements in this formula can be obtained, leading to the following discrete choice analysis-based formulation of the metanormative framework: $\hat{V}(a_i) = \sum_t^T \left[\hat{P}(t) \cdot \hat{V}_t(a_i)\right]$, which gives the estimated utility of action i, taking into account that people in different classes ascribe different utilities to that action.

[2]The class membership probability is computed by a logit function which ensures that $\sum_t^T P(t) = 1$.

6.4. IMPLEMENTATION OF THE DISCRETE CHOICE ANALYSIS-BASED FORMULATION OF THE METANORMATIVE FRAMEWORK

To guide the process of Discrete Choice Analysis for the implementation of the metanormative framework, we use a two-pronged approach with long pedigree in the applied choice literature. It comprises the specification of the utility function and the collection of data to allow for a statistically rigorous estimation of the utilities of alternatives. While the formulations presented in the previous section are general, we will now interpret them in light of the data we use for our empirical proof of concept. These data resulted from a choice experiment that took place in 2017 which entailed a public consultation for a massive national transport infrastructure scheme among car commuters in The Netherlands. The scheme was specified in terms of its consequences on a number of dimensions relative to the *status quo* [184].

For the utility specification, we build on the seminal work in consumer theory by Lancaster, who postulated that utility is derived not from goods but rather from properties or characteristics of the goods [455], and we follow the utilitarianism view on ethics by considering the utility of an action a_i to be a linear function of the action's consequences: $V_t(a_i) = \beta_{it} + \sum_m \beta_{mt} \cdot x_{mi}$. Here, $V_t(a_i)$ is the utility ascribed by a member of class t to action a_i; x_{mi} is the mth consequence of action a_i; β_{mt} is the weight attached by a member of class t to that consequence; and β_{it} is the remaining – i.e., not associated with any observed consequence – utility of the action. Importantly, a vector of weights or β is considered to represent a multidimensional moral theory. Further research is to be explored using more sophisticated behavioral representations such as Random Regret Minimization or Taboo models instead of a linear utility function [456, 457], as such allowing for a richer representation of various moral preference structures and decision rules.

Concerning the data collection, it is noted that choice experiments are widely used in the applied choice domain [458]. The key point in these experiments is to systematically vary the consequences (attributes) of actions (alternatives) and construct choice tasks by combining different hypothetical actions. Choices between these, or between one action and an opt out (*status quo* action), then give maximum information about preferences which allow for statistically efficient estimation of the weights β_{mt} for each consequence β_{it}. The data which we use in our empirical proof of concept resulted from a full factorial choice experiment [3] in which 99 respondents, composing a roughly representative sample of the Dutch regular car commuters, were presented with a series of binary choice situations with a clear *status quo* (current situation) and an alternative (a proposed infrastructure investment scheme). The infrastructure investment scheme was specified in terms of its positive or negative consequences with respect to vehicle ownership tax, travel time for the average commute trip, as well as the number of seriously injured in traffic, and the number of traffic fatalities[4]. The final experiment resulted in $99 \times 16 = 1584$ choice observations (Tables C.1 and C.2 in Appendix C) [184].

The utility of a particular infrastructure transport policy is written as $V_j = \sum_m \beta_m \cdot x_{jm} = \beta_{tax} \cdot tax_j + B_{time} \cdot time_j + \beta_{inj} \cdot inj_j + \beta_{fat} \cdot fat_j$. The utility of the opt out (status quo)

[3] A description of the full set of actions $i = 1$ to $i = 16$ is found in the Supplementary Information.

[4] The consequences were effect-coded as [−1] for a decrease in the level of attributes and [1] for an increase in the level of attributes (Table C.1 in Appendix C). For example [-1] on vehicle ownership tax means a decrease of 300 euros in the vehicle ownership tax per year.

option is merely defined in terms of a so-called alternative specific constant (ASC) which represents a generic preferences for or against the status quo versus an infrastructure investment scheme. In the sampled population, different classes are found featuring a different vector of β, which implies a different weighing of the consequences, thus defining the different trade-offs the members of a class are willing to accept. Because such trade-offs involve the well-being of humans, we postulate that we can use the vector of β to infer the morality of the classes in this particular context.

It is relevant to note that, while our operationalization is well aligned with a consequentialist view on ethics, in this research we refrain from relating this morality vector to particular moral theories, such as deontology or consequentialism and their ramifications [459]. Rather, we explore the subtle differences in contextual moral preferences that characterize different classes and are captured in the vector of β that defines each class. The empirical work to implement the discrete choice analysis-based metanormative theory is described below, followed by a proof of concept that allows us to investigate whether the policy choices made by an AI System, based on the conflicting input of different sized moral classes, would differ from the same choices made by an AI System that overlooks such differences (Tables 6.1 and 6.2).

Table 6.1: Attributes in choice experiment.

ATTRIBUTES	INCREASE/DECREASE
Vehicle ownership tax (Euros)	300 per year
Travel time (Minutes)	20 per working day
Non-fatal traffic injuries	100 per year
Traffic fatalities	5 per year

Table 6.2: Example of a choice task in the choice experiment.

	Proposed Transport Policy
Vehicle ownership tax (per year, for each car owner including yourself)	300 euro **less** tax
Travel time (per working day, for each car commuter including yourself)	20 minutes **less** travel time
Number of seriously injured in traffic (per year)	100 seriously injured **more**
Number of traffic fatalities (per year)	5 traffic fatalities **more**
YOUR CHOICE	☐ I support the proposed policy ☐ I oppose the proposed policy

6.5. EMPIRICAL ANALYSIS

We describe the empirical approach that is employed to capture and investigate the relevance of moral uncertainty in AI decision-making by first elaborating on the model estimation, followed by a brief characterization of the classes in the model, and an inspection of the utility of actions per class.

6.5.1. MODEL ESTIMATION

To decide on the optimal number of latent classes, consecutive models with one through four classes were estimated and compared on the dataset that resulted from the choice experiment described above. In general, the decision to select a certain number of latent classes is a trade-off between model fit (in terms of the log-likelihood) and parsimony (in terms of the number of classes/parameters) and interpretability.

Typically, such a decision is therefore guided by an information criterion, which weighs both model fit and parsimonity. In the context of latent class modeling, the Bayesian Information Criterion (BIC) criterion has been shown to perform well [224]. The BIC is a fit criteria for model selection that measures the trade-off between model fit and complexity of the model [460]. The equation used to calculate this criterion is $BIC = (\ln N) \cdot k - 2(\ln L)$ where N is the number of recorded measurements (e.g. choices), k is the number of estimated parameters, and L is the maximum value of the likelihood function for the model. A lower BIC indicates a better model.

In the present application, this statistic indicated that the optimal solution is one with four or more classes, which would be too many to interpret meaningfully (Table 6.3). A straightforward and practical alternative to the BIC is to compute the percentage increase in the log-likelihood of each model compared to the baseline one-class model. This measure reveals that after three classes there is no substantial increase in the relative fit of the model (LL increase > 4%).

Table 6.3: BIC and Log-Likelihood function models 1-4.

Model	BIC	Log-Likelihood Function
Model 1	1479.290	-721.226
Model 2	1400.273	-659.614
Model 3	1369.509	-622.129
Model 4	1360.966	-595.754

6.5.2. SIZE AND FEATURES OF THE CLASSES

The classes that compose the three-class model have different sizes and defining features. For the assessment of the size of the classes, we recall the discrete choice analysis-based formulation of the metanormative framework that was introduced above: $\hat{V}(a_i) = \sum_t^T \left[\hat{P}(t) \cdot \hat{V}_t(a_i)\right]$. The estimate of the probability $\hat{P}(t)$ that a randomly sampled individual belongs to class t equals the relative share of the population that belongs to that class (Table 6.4). A vector of β estimated in the empirical process is associated to each class allowing us to understand the key defining features and the subtle differences in moral preferences that characterize the different classes (Table 6.5). We provide below an interpretation of the

features of each class, along with the full estimation and interpretation of the parameters, that define each class (Table 6.5).

Class 1: *Financially-driven* The first class is the smallest segment in the three-class model (≈ 14%). Its members care only about lowering vehicle ownership taxes. We therefore infer that members of this class are financially-driven. All parameters have negative signs but only vehicle ownership tax is statistically significant, which means that for the members of this class the utility of a policy decreases if it features an increase in the vehicle ownership tax.

Class 2: *Want-it-all's* The second class is the largest segment in the three-class model (≈ 65%). Its members show a generic disposition against policies and a preference for lower vehicle ownership taxes, lower travel time, lower number of seriously injured in traffic, and lower number of traffic fatalities; they are especially concerned with lowering the number of seriously injured in traffic. We infer taht members of this class are maximizers that want it all and believe that changes should only occur if a substantial improvement is secured, specifically in terms of road safety. All parameters are statistically significant and, with the exception of the alternative specific constant (ASC) for the status quo option, have negative signs. This means that for the members of this class the utility of a policy decreases if it features an increase in the vehicle ownership tax, travel time, number of seriously injured in traffic, or number of traffic fatalities.

Class 3: *Efficient* The third class accounts for ≈ 22% of the sampled population. Members of this class care mostly about low travel time and therefore we infer that they are (time-) efficient. With the exception of the ASC for the status quo, all parameters have negative signs and are all statistically significant, except for the traffic fatalities parameter which has low significance. This means that the members of this class show a disposition against policies and consider that the utility of a policy decreases if it features an increase in the vehicle ownership tax, travel time or number of seriously injured in traffic.

Table 6.4: Class membership probability for classes in three-class model.

Classes	Class membership probability
Class 1: Financially-driven	0.14
Class 2: Want-it-all's	0.65
Class 3: Efficient	0.22

Table 6.5: Estimated parameters in the classes of three-class model.

Name	Value	Std err	t-test	p-value
Class 1: Financially-driven				
ACS Oppose	-0.519	0.359	-1.44	0.15
BETA Fat	-0.561	0.298	-1.88	0.06
BETA Inj	-0.209	0.288	-0.73	0.47
BETA Tax	-2.56	0.339	-7.54	0
BETA Time	-0.119	0.253	-0.47	0.64
Class 2: Want-it-all's				
ACS Oppose	1.52	0.136	11.16	0
BETA Fat	-1.41	0.14	-10.09	0
BETA Inj	-1.92	0.169	-11.32	0
BETA Tax	-0.967	0.117	-8.25	0
BETA Time	-0.328	0.111	-2.97	0
Class 3: Efficient				
ACS Oppose	1.24	0.222	5.59	0
BETA Fat	-0.36	0.189	-1.9	0.06
BETA Inj	-0.745	0.186	-4	0
BETA Tax	-1.02	0.189	-5.38	0
BETA Time	-1.72	0.264	-6.52	0
s2	1.54	0.333	4.62	0
s3	0.442	0.425	1.04	0.3

6.5.3. UTILITY OF ACTIONS

We have so far determined the number of latent classes in the model and provided a brief description of each class. Now we proceed to inspect the utility and rank of actions (i.e. policies) in the three-class model and in each of its classes. We recall once again the discrete choice analysis-based formulation of the metanormative framework that was introduced above $\hat{V}(a_i) = \sum_t^T \left[\hat{P}(t) \cdot \hat{V}_t(a_i)\right]$ to accentuate that the estimate of the utilities $\hat{V}_t(a_i)$ of policies in each class is given by $V_j = \sum_m \beta_m \cdot x_{jm} = \beta_{tax} \cdot tax_j + B_{time} \cdot time_j + \beta_{inj} \cdot inj_j + \beta_{fat} \cdot fat_j$. Such class-specific utilities are subsequently multiplied by the class membership probability of each class $\hat{P}(t)$ and summed over classes for the purpose of estimating the utility of policies $\hat{V}(a_i)$. To facilitate the comprehension about the utilities in the three-class model, we compare it with the baseline one-class model (Table 6.6) [5]. This comparison will take a new meaning in Section 6.6, as we conceptualize a society in which AI Systems equipped with one-class and three-class rules make policy decisions on behalf of society.

[5]The parameters for the one-class model can be found in the Appendix C (Table C.2).A description of the full set of actions $i = 1$ to $i = 16$ effect-coded as [−1] for a decrease in the level of attributes and [1] for an increase in the level of attributes can also be found in the Appendix C.

Unsurprisingly, the policy with highest utility in both models is $i = 1$ which entails lower vehicle ownership taxes, lower travel time, lower number of injuries, and lower number of fatalities; followed by $i = 12$ which entails higher travel time but lower vehicle ownership taxes, lower number of injuries, and lower number of fatalities; and by $i = 2$ which entails more traffic fatalities but lower vehicle ownership tax, lower travel time, and lower number of injuries (Table 6.6). It is clear that the utility of the policies in the three-class model is highly influenced by the preferences of the *Want-it-all's*, which make up the largest class in the model.

Table 6.6: Utility of policies $\hat{V}(a_i)$ in one-class model, three-class model, and class-specific utility of policies $\hat{V}_t(a_i)$ in classes [1-3] of three-class model.

i	One-class Model	Three-class Model	Class 1	Class 2	Class 3
1	4.327	5.471	2.93	6.145	5.085
2	2.743	3.339	1.808	3.325	4.365
3	0.523	0.479	1.39	-0.515	2.875
4	-0.517	-0.719	1.152	-1.171	-0.565
5	-2.473	-3.117	-3.968	-3.105	-2.605
6	2.371	3.073	-2.19	4.211	3.045
7	1.331	1.875	-2.428	3.555	-0.395
8	-0.889	-0.984	-2.846	-0.285	-1.885
9	1.703	2.141	1.57	2.669	0.925
10	0.151	0.214	-2.608	0.371	1.555
11	2.107	2.612	2.512	2.305	3.595
12	3.287	4.273	2.692	5.489	1.645
13	1.067	1.414	2.274	1.649	0.155
14	0.787	0.941	-3.312	1.391	2.325
15	-1.433	-1.919	-3.73	-2.449	0.835
16	-0.253	-0.257	-3.55	0.735	-1.115

In order to measure the rank correlation of the utility of policies in the baseline one-class model, the three-class model, and the different classes within the three-class model a Kendall Tau test was used (Table 6.7). We observe that the baseline one-class model and the three-class model have the same ranking of policies (correlation = 1.0), which raises questions about the relevance of taking into account moral uncertainty for the purpose of ranking actions. This will be addressed later (Section 6) as we will randomly generate thousands of policies and evaluate the implications of moral uncertainty across simulated cases. We also used the Kendall Tau coefficient test the rank correlation of the utility of policies among the three classes in the three-class model. The ranking of the policies by the members of the class 1 (*Financially-driven*) and class 2 (*Want it all's*) show a low correlation (correlation = 0.533) similarly to the rankings of the policies by (*Financially-driven*) and class 3 (*Efficient*) (correlation = 0.483) (Table 6.8).

Looking specifically at the classes within the three-class model, we observe that, similarly

Table 6.7: Kendall Tau b-test rank correlation of policies between one-class model and three-class model; one-class model and classes [1-3] of three-class model; and three-class model and classes [1-3] of three-class model .

	Three-class model	**Class 1**	**Class 2**	**Class 3**
One-class model	1.000	0.700	0.833	0.790
Three class model	(N/A)	0.688	0.969	0.764

Table 6.8: Kendall Tau b-test rank correlation of policies between classes [1-3] in the three-class model.

	Class 1	**Class 2**	**Class 3**
Class 1	N/A	0.533	0.483
Class 2	0.533	N/A	0.450
Class 3	0.483	0.450	N/A

to what was reported above concerning the baseline one-class model and the three-class model, the policy with highest utility in all classes is $i = 1$ which entails lower vehicle ownership taxes, lower travel time, lower number of injuries, and lower number of fatalities. On the other hand, the policy with lowest utility in all three classes is $i = 5$ which entails higher vehicle ownership taxes, higher travel time, higher number of injuries, and higher number of fatalities. Moreover, by computing the standard deviation of the ranks for each policy in the different classes of the three-class model, we determine the discrepancy in rankings among the classes (Table 6.9). Three policies registered high discrepancies in rankings: *i=3*, *i=7*, and *i=13*. We remark that policy *i=3*, which entails lower vehicle ownership tax and travel time, and higher injuries and traffic fatalities [6], ranks substantially higher among the *Efficient* and *Financially-driven* compared with the *Want it all's*. Policy *i=7* entails higher vehicle ownership tax and travel time, and lower injuries and traffic fatalities ranks much higher among *Want it all's* when compared to the *Efficient* and *Financially-driven*. And finally policy *i=13*, which entails lower vehicle ownership tax, higher travel time and also higher number of traffic injuries, and lower fatalities, ranks substantially higher among the *Financially-driven* when compared to the *Efficient* and in a lesser extent to the *Want it all's*.

[6] We refer once again to Table C.1 in Appendix C.

Table 6.9: Standard deviation of the ranks for each policy in classes [1-3] in three-model class.

i	Standard deviation
1	0
2	1.41
3	3.40
4	2.62
5	0
6	2.62
7	3.40
8	1.41
9	1.41
10	1.41
11	1.89
12	2.36
13	2.87
14	2.87
15	2.36
16	1.89

We have shown that there is variance in the utility assigned to particular policies by members of different classes. These discrepancies seem to arise when policies involve trade-offs as opposed to policies that merely have desirable or undesirable outcomes, which is relevant given that policies in general tend to involve trade-offs. Using the same dataset we will now provide a proof of concept to further explore the relevance of moral uncertainty in the context of AI.

6.6. PROOF OF CONCEPT: AI SYSTEMS MAKE POLICY CHOICES ON BEHALF OF SOCIETIES

To illustrate the discrete choice analysis-based formulation of the metanormative framework, we capitalize on the fact that the AI field is traditionally lenient to remarkable thoughts. We therefore conceptualize AI Systems that make policy choices on behalf of a society. And we further investigate whether differences in policy choices arise as a result of accounting for moral uncertainty.

Building on the work outlined in previous sections, we consider an AI System equipped with a one-class rule and an AI System equipped with a three-class rule which factors in moral uncertainty. To make a larger action space available for the AI Systems, we randomly generated 5000 sets with 2 policies in each set by allowing the consequences of the policies, i.e. x_{jm} in the formulation of the utility of a infrastructure transport policy $V_j = \sum_m \beta_m \cdot x_{jm} = \beta_{tax} \cdot tax_j + B_{time} \cdot time_j + \beta_{inj} \cdot inj_j + \beta_{fat} \cdot fat_j$, to take on random

values within the interval $[-1,1]$ instead of taking only the extreme values as in the original dataset (Table C.1 in Appendix C) [7].

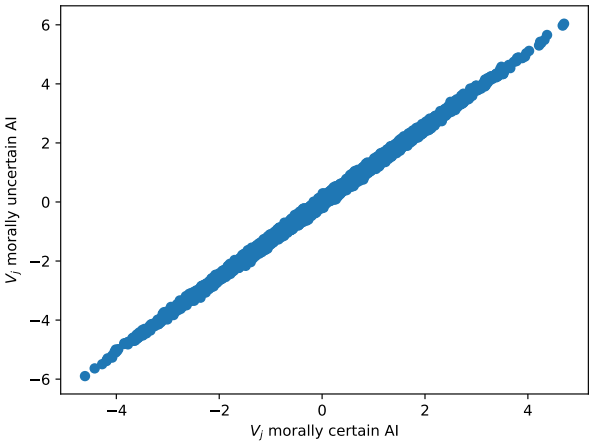


Figure 6.1: Scatter plot of utilities of baseline and three-class morally uncertain model for 5000 randomly drawn policies.

The utility of each policy was estimated through the formulations introduced above for both the one-class baseline model (for the morally certain AI System) and for the three-class model (the morally uncertain AI System). The comparative value of the utilities of policies is interpreted as an indication of which policy is favored by each model and accordingly by the corresponding AI System. Although the utilities assigned to policies by the morally certain and the morally uncertain AI seem to be similar in most cases (Figure 6.1), we observed 85 cases in which there was disagreement between the two models regarding the choice of the policy (Figure 6.2).

In such instances of disagreement, the policy decisions of the morally uncertain AI system equipped with the three-class rule would contrast with the decisions of the morally certain AI equipped with the baseline one-class rule. Although the number of cases of disagreement is not large in this particular example it still allows us to hint at the potential relevance of capturing moral uncertainty. To visualize the contrasting policy decisions, we plotted the difference, for each of the 85 combinations of policies, between the utility of the first policy and that of the second policy for the morally certain AI System equipped with the one-class model (horizontal axis) and for the morally uncertain AI System equipped with the three-class model (vertical axis) (Figure 6.2).

Out of the 85 cases mentioned above we selected three sets of policies that registered relatively high discrepancies between the utility values estimated for the baseline (morally certain) AI and those estimated for the morally uncertainty AI. The first set of policies that we selected featured a *time efficient and safe for injuries* policy favored by the baseline AI and a *safe but expensive and time inefficient* policy favored by the morally uncertain AI

[7] For example [-0.5] on vehicle ownership tax means a decrease of 150 euros in the vehicle ownership tax per year

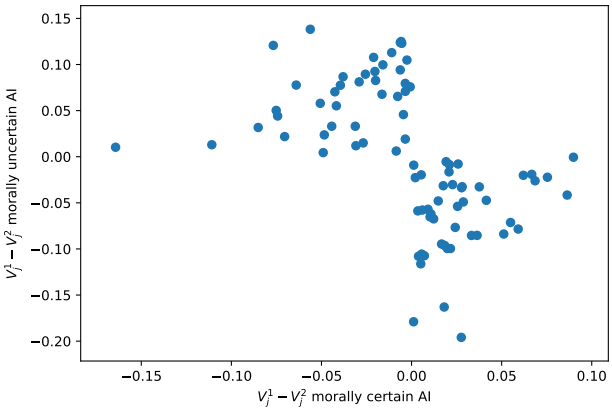


Figure 6.2: Policies that caused disagreement among baseline and three-class morally uncertain model.

(Figure 6.3).

Another set of policies featured a *time efficient but unsafe for fatalities* policy favored by the base model and a *safe but time inefficient* policy favored by the morally uncertain model (Figure 6.4).

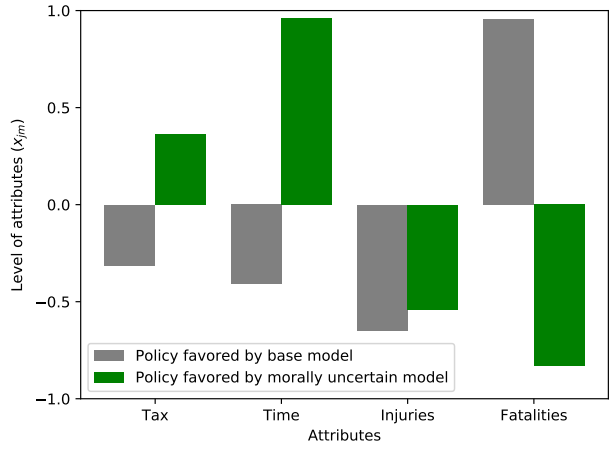


Figure 6.4: Set of policies: *time efficient but unsafe for fatalities* policy and *safe but time inefficient* policy.

Finally the third set of policies registering high discrepancies in utility value among the two competing models features a *time efficient but unsafe* policy favored by the base model and a *safe but expensive* policy favored by the morally uncertain model (Figure 6.5).

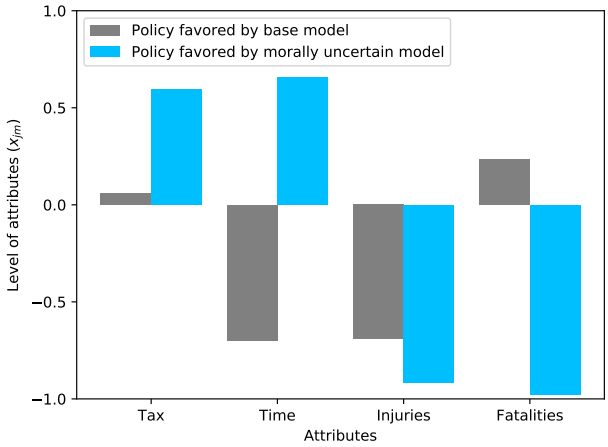


Figure 6.3: Set of policies: *time efficient and safe for injuries* policy and *safe but expensive and time inefficient* policy.

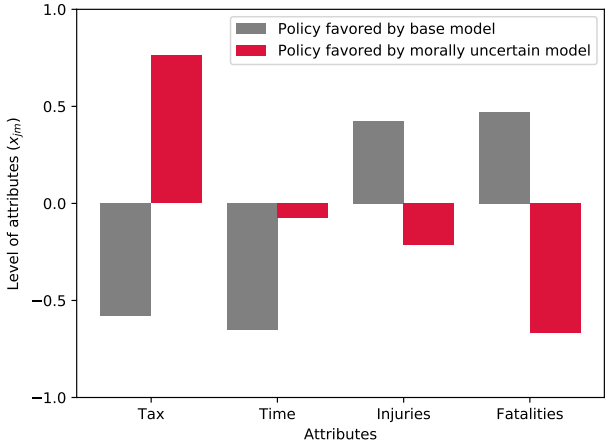


Figure 6.5: Set of policies: *time efficient but unsafe* policy and *safe but expensive* policy.

These discrepancies in policies chosen by the morally certain AI and the morally uncertain AI emphasize the relevance of studying moral uncertainty and capturing it in AI systems.

6.7. CONCLUSION AND DISCUSSION

As AI Systems become increasingly autonomous, they are expected to engage in complex moral decision-making processes. For the purpose of guidance of such processes, theo-

retical and empirical solutions within the controversial domain of Machine Ethics have been sought. In this research we integrate both theoretical and empirical lines of thought to address the matters of moral reasoning in AI Systems in a pragmatic yet statistically rigorous way that is firmly connected to theoretical considerations related to normative uncertainty in AI systems.

Our approach is built on the theoretical notion that moral uncertainty is paramount to the endeavor of implementing moral reasoning in AI Systems. More specifically, it employs the metanormative framework for decision-making under moral uncertainty, as construed by William MacAskill and Kyle Bogosian [398, 429], and re-conceptualizes it as a latent class discrete choice model. We assume that a number of classes featuring different preferences exist within a population where each class is internally relatively homogeneous in terms of its behaviors and preferences [450]. By codifying the moral preferences and decision rules of different classes, and aggregating them across the population, we are able to obtain a moral representation for the AI System: its resulting normative uncertainty is embedded in the form of an empirical model of the moral heterogeneity of the society it represents.

In the empirical installment of our approach we specify a multi-dimensional utility function which represents a moral theory or set of moral preferences (i.e., weights attached to different criteria), and we allow this vector of weights to vary across classes. The final ranking of the actions available to the AI System is provided by computing the class membership weighted average of the utility of each action in each class. Importantly, our approach does not involve the analyst a priori selecting classes and class sizes in the population, they rather emerge - just like the class-specific weights assigned to each criteria - in the process of estimating the choice model from observed choice data.

The discrete choice analysis-based formulation of the metanormative framework is theory-rooted and practical, as it captures moral uncertainty through a small set of latent classes, thus avoiding runtime issues which are common in applications that aim to capture the full level of individual-to-individual heterogeneity in the population.

For the purpose of illustrating our approach we conceptualize a society in which AI Systems are in charge of making policy choices. In the proof of concept two AI systems make policy choices on behalf of a society, but while one of the systems uses a baseline morally certain model the other uses a morally uncertain model. Specifically, we used our approach in a dataset that resulted from a choice experiment that took place in 2017 which entailed a consideration for a massive national transport infrastructure scheme among car commuters in The Netherlands, having implications on morally salient dimensions such as the number of road fatalities. It was observed that there are cases in which the two AI Systems disagree about the policy to be chosen, which we believe is an indication about the relevance of moral uncertainty.

We are aware that our finding that a morally uncertain AI might in some cases decide differently than a morally certain AI not only validates the notion that moral uncertainty is a topic worthy of further investigation by the AI community, but that it also generates another question: in cases where the two AIs would make different decisions, which AI should prevail? This question is not one with a clear-cut answer, but the following observations and considerations could help address this matter.

First, it is important to revisit the starting point of this research: our aim was to present

an approach to capture moral uncertainty in AI that a) has a firm theoretical foundation, and b) is empirically and practically feasible. To achieve both aims simultaneously, we proposed a latent class approach, which is a compromise between not acknowledging moral uncertainty at all (i.e., estimating one set of moral preferences to represent an entire society's morality) and taking into account subtle moral differences between each and every individual in society. By allowing a small number of latent classes with distinct morality to emerge in a process of econometric model estimation, we connect to the theoretical framework of normative uncertainty, and avoid run-time issues which plague the estimation and application of individual-level models. Building on this argument, we believe that, in general terms, the decisions made by a morally uncertain AI (equipped with a latent class choice model of morality) should be preferred to the decisions made by an AI that is morally certain.

Whether or not the decisions made by a morally uncertain AI equipped with a limited number of latent moral classes are to be preferred over those made by an AI that tries to embed the morality of each individual in society (supposing that this would be feasible), is another matter.

Here, we propose to use the notion of Occam's razor, which puts a premium on the most simple explanation behind empirical observations. In statistics, this generic scientific notion is operationalized in metrics such as the adjusted rho-square, the Bayesian Information Criterion and others, which penalize a model for the number of parameters it uses to explain data. In the field of machine learning, this relates to the notion of regularization, which prevents artificial neural networks from over-fitting training data-sets. Such statistical tools and metrics offer a formal, theory-rooted approach to select one model (or: AI system) over another. For example, in our case, allowing for a small number of latent moral classes clearly led to a better explanation for the observed choice data than a model that attempts to embed all society's preferences in one utility function, also after correcting for the increased number of parameters.

However, further increasing the number of latent classes beyond a handful inevitably leads to increases in the number of parameters that no longer are offset by increases in explanatory power. The same holds for morally uncertain AI Systems that aim to capture the differences in morality between each and every individual: the resulting model, besides being difficult to handle in real time decision contexts, will most likely be statistically inferior to more parsimonious models that attempt to cluster individuals that are relatively speaking like-minded in terms of their morality. In sum, by employing statistical model selection techniques that appropriately penalize for the number of parameters used, helps the designer of the AI choose the optimal level of heterogeneity (uncertainty) to embed in the AI.

The novelties in this research are the idea that discrete choice analysis can be used to codify human morality and, as such, provides a tool to embed morality in AI systems; moral uncertainty can be operationalized by re-conceptualizing the metanormative framework of normative uncertainty through latent class choice models; and also the empirical illustration of the concept of moral uncertainty. We acknowledge that our re-conceptualization fails to take into account the richness and subtleties of the work developed originally by MacAskill[429, 430] yet opening an avenue for further research that accommodates its various extensions. Additionally, instead of using a linear utility function, as it was the case

in this research, other utility functions such as Random Regret Minimization [457], taboo trade off aversion, or lexicographic choice may be explored, as these indirectly refer to different moral theories. Finally, through the proof of concept this research also opens avenues for further research on the meaning and practical implications of moral uncertainty in artificial decision-making.

7

CONCLUSION

Earlier in this doctoral dissertation, it was asserted that the renewed interest in AI is a moment of truth for AI Ethics. Normative discussions about AI have run parallel to its development throughout the decades. Still, only recently have these discussions received an overwhelming amount of attention from the scientific, governmental, non-governmental, and media communities. No longer confined to a small space of moral debate, AI Ethics has now the opportunity to mature as an applied branch of Ethics.

Despite the richness of AI Ethics work, a growing body of the literature has declared that it is failing to realize its normative endeavors. It is hypothesized that the current struggles of AI Ethics may be rooted in the challenges that have emerged in the data-driven paradigm in AI. In this larger moral space, AI Ethics faces important challenges related to normative urgency, multi-purpose nature of AI technology, and multitude of AI stakeholders. This dissertation builds on the premise that empirical information is valuable for AI Ethics to address these challenges and realize its normative mandate. A collection of empirical studies is presented along the spectrum of empirical ethics and morally relevant social sciences, thus making a case for incorporating more empirical research into AI Ethics.

In this dissertation *Ethical AI* is conceptualized in three different dimensions: (i) data analysis tool in morality; (ii) system in morally charged context; and (iii) artificial moral agent. Each dimension is operationalized through empirical methods from the social sciences toolkit. There is an expectation that this research contributes to the reflective development process, which assists the communities operating in the AI space to engage in a critical reflection about AI.

The findings reported in this dissertation are summarized (section 7.1). Subsequently, scientific (section 7.2) and policy (section 7.3) implications of this research are formulated. The dissertation concludes with a final reflection (section 7.4).

7.1. MAIN FINDINGS OF STUDIES IN THE DIFFERENT DIMENSIONS OF ETHICAL AI

7.1.1. AI AS A DATA ANALYSIS TOOL IN MORALITY

RG1: TO ILLUSTRATE HOW AI-TECHNIQUES CAN BE USED TO INVESTIGATE THE RELATION BETWEEN MORALITY AND BEHAVIOR

The study presented in the first dimension of *Ethical AI* (**Data analysis tool in morality**) addressed a research goal of using unsupervised learning algorithms to gain insights into morality. This research investigated the relationship between the moral foundations posited by the Moral Foundations Theory (MFT) and morally relevant human behavior. The case study was gender-based violence legal cases decided in Portuguese (appeal) courts. The moral foundations were mapped in the morally charged legal cases and the corresponding legal outcomes were conceptualized as morally relevant behavior. A unsupervised learning algorithm (latent class cluster analysis) was used to identify complex patterns in the data and reduce such patterns to a distinct number of clusters. This study contributes to MFT research by further exploring the association between moral foundations and behavior, thus making a case for the use of these algorithms in the domain of morality.

- There is an association between moral foundations in gender-based violence narratives of legal cases decided in Portuguese courts and corresponding legal outcomes.
- There is a dominance of individualizing moral foundations in gender-based violence narratives of legal cases decided in Portuguese courts.
- The care foundation is associated with clusters in which the offender is often convicted to prison time.
- The fairness foundation is associated with clusters in which the offenders spend little or no time in prison due to procedural decisions related for instance to suspended sentences or re-trial.
- When adjusting the level of courts at the intermediate appeal level, the sanctity foundation is associated with the cluster featuring sexual offenses.

7.1.2. AI AS A SYSTEM OPERATING IN MORALLY CHARGED CONTEXTS

The studies presented in the second dimension of *Ethical AI* (**System operating in morally charged contexts**) addressed key research questions about AI in Transportation and Healthcare.

RQ1: WHAT ARE THE ETHICAL ISSUES IN FOCUS BY THE AUTONOMOUS VEHICLES INDUSTRY?

This study investigated industry narratives about the ethics of autonomous driving. A systematic review of both scientific and industry literature shed light on the key ethical issues associated with AVs. From the scientific literature, critical yet practical questions related to safety, accountability, and human oversight were raised, for which answers were searched in the industry reports. The findings reported in this study are important insights expected to improve communication between the different stakeholders in autonomous driving.

- Industry and academia look at the ethics of AV technology through rather different lenses.
- While the scientific literature has been largely preoccupied with deep considerations of abstract moral dilemmas (trolley problem), industry reports adopt a much more pragmatic, technology-infused and perhaps overly optimistic narrative when discussing the potential of so-called edge cases where accidents cannot be avoided and loss of life and damage need to be minimized.
- Given the plethora of ethical issues addressed in the reports, AV companies seem to be aware of the ethics of autonomous driving technology.
- Both scientific and industry literature prioritize safety and cybersecurity and agree that AVs will not eliminate the risk of accidents.
- The scientific literature on the ethics of AVs is dominated by discussions about the trolley problem.
- Moral dilemmas resembling trolley cases are not addressed in industry reports but there are nuanced allusions that unravel underlying concerns about these extreme traffic situations.
- Autonomous driving companies have different approaches with respect to the authority of remote operators.
- Companies seem invested in a lowest liability risk design strategy relying on rules and regulations, expedite investigations, and crash/ collision avoidance algorithms.

RQ2: WHAT ARE THE PERSPECTIVES OF MEDICAL PRACTITIONERS ABOUT HEALTH AI?
This study surveyed medical practitioners based in different countries on a broad scope of ethical issues regarding Health AI. Four main perspectives were identified. By unraveling these different perspectives, this study allowed a better understanding of the perceptions of practitioners on the ethical risks of AI in the healthcare setting.

- Four main perspectives were identified about Health AI.
- P1. *AI is a helpful tool: Let physicians do what they were trained for*
  In this perspective there is an overall positive outlook about the implementation of AI technology in healthcare. AI is regarded as a helpful tool that will allow doctors to have the time to focus on top-of-license skill sets and activities. However, medical doctors must remain in charge of the medical decision process and participate in the technology design process.
- P2. *Rules & Regulations are crucial: Private companies only think about money!*
  In this perspective there is a clear negative outlook about AI technology and distrust in private health companies. There is a sentiment that the tech industry is not well aligned with core healthcare values, has little knowledge about Medicine, and poses a risk for monopolistic behavior
- P3. *Ethics is enough: Private companies can be trusted*
  In this perspective there is an overall positive outlook about AI companies. It is not undesirable that these companies start operating in the health space. Moreover, there are no major concerns about the risk for monopolistic behavior.

- P4. *Explainable AI tools: Learning is necessary and inevitable!*
  In this perspective, Explainability is a crucial value: in order to reap the benefits of Health AI, medical doctors must understand and lead the technological process.
- Most areas of healthcare can benefit from AI.
- AI must be aligned with core bioethical principles.
- Medical practitioners must participate in the AI design process.

7.1.3. AI AS AN ARTIFICIAL MORAL AGENT

RQ3: WHAT ARE THE PERSPECTIVES OF AI ETHICS SCHOLARS ABOUT ARTIFICIAL MORAL AGENTS?

This study surveyed AI Ethics scholars on controversial issues about artificial morality ((i) development of AMAs; (ii) design of AMAs; (iii) moral agency of AMAs; and (iv) future projections of AMAs). Five main perspectives were identified, thus establishing an informed debate on artificial morality, where contrasting views and agreements are disclosed and appreciated. This study improves the communication of normative work on artificial morality to different communities operating in the AI space.

- Five main perspectives were identified about artificial morality.
- P1. *Machine Ethics: The Way Forward*
  AMAs are unavoidable and may be necessary for technological progress. Moral sureness and a mixed design approach are essential, but free will is not. More than simple tools, AMAs will advance our understanding of morality.
- P2. *Ethical Verification: Safe & Sufficient*
  AMAs will not replace humans in ambiguous moral situations, given that ethics and human moral agency are not algorithmic. Transparency, accountability, and predictability lead to sufficiently ethical machines.
- P3. *Morally Uncertain Machines: Human values to Avoid Moral Dystopia*
  AMAs must be morally uncertain and hold human values. Otherwise, they pose an existential threat. Simply prohibiting unethical behavior, as well as implementing external checks and balances, is not enough.
- P4. *Human Exceptionalism: Machines Cannot Moralise*
  AMAs lack moral agency as they do not have free will nor understanding to make moral assessments as humans do. Logical machines will not be better moralizers or our moral teachers.
- P5. *Machine Objectivism: Machines as Superior Moral Agents*
  AMAs prevent human harm. Through logic and context-specificity, they are better moral reasoners and educators. Free will and conceptual understanding are not required for moral agency.
- A potential source of the differing perspectives is the failure of Machine Ethics to be widely observed or explored as an applied branch of Ethics and more than a futuristic end.

- AI systems working in morally salient contexts cannot be avoided and, as such, some degree of moral competence ought to be demonstrated.
- There is a long way to go before AMAs replace human beings in difficult moral situations.

RG2: HOW TO USE AN EMPIRICAL APPROACH TO CAPTURE MORAL UNCERTAINTY FOR THE DESIGN OF ARTIFICIAL MORAL AGENTS

This study addressed a research goal of using an empirical approach to capture moral uncertainty in the design of AMAs. The main premise is that moral uncertainty is paramount to the endeavor of implementing moral reasoning in AI Systems (AMAs). A metanormative framework for decision-making under moral uncertainty was operationalized through a discrete choice model. To illustrate this approach, a society in which AI systems are in charge of making policy choices was conceptualized. In the proof of concept, two AI systems make policy choices on behalf of society. However, while one of the systems uses a baseline morally certain model, the other uses a morally uncertain model. It was observed that there are cases in which the two AI Systems disagree about the policy to be chosen, thus signaling the relevance of moral uncertainty.

- Discrete choice analysis can be used to codify human morality.
- Latent class choice models can be used to operationalize moral uncertainty.
- The empirical illustration of moral uncertainty opens avenues for further research on the meaning and practical implications of moral uncertainty in artificial decision-making.

The findings reported above with respect to the various dimensions of *Ethical AI* are expected to assist the communities operating in the AI space engaging in critical reflection about AI. In addition to these findings, overarching conclusions about this dissertation are formulated, which may be relevant for further AI Ethics research.

7.2. CONTRIBUTIONS OF RESEARCH

7.2.1. CONTRIBUTIONS FOR THE REFLECTIVE DEVELOPMENT ABOUT AI

This dissertation was built around the idea that **empirical information is valuable** for AI Ethics to meet its normative goal of providing guidance that allows society to capitalize on the innovation benefits of AI while minimizing its risks.

It was premised that empirical information would provide (i) much needed **systematization** to address the challenge of normative urgency; (ii) clear **perception of the ethical risks** to address the challenge of multi-purpose technology; and (iii) improved **communication, insights, and operationalization** of theoretical concepts to address the challenge of multitude of stakeholders operating in the AI space.

The research featured in this dissertation has met these premises. The empirical approach in AI Ethics allowed the systematization of information, thus facilitating the normative guidance of AI. Moreover, this research, namely the studies on AVs (chapter 3) and Health AI (chapter 4), contextualizes the risks, conflicting rights and interests, and social preferences associated with these particular AI Systems in different societal domains. Finally, by operationalizing theoretical and abstract concepts, this research, namely the studies on artificial

morality (chapter 5) and moral uncertainty (chapter 6), improves the communication with other stakeholders and communities.

There is an expectation that this empirical endeavor will contribute to a **reflective development about AI**. The studies featured in this dissertation provide important systematic and operational insights into the normative issues of AI. These insights should contribute to a critical development of AI. However, as mentioned earlier, there is no conclusive evidence or measurement of such contribution. This is a common limitation in exploratory research, which is aggravated by the complex nature of this subject matter.

7.2.2. SCIENTIFIC CONTRIBUTIONS

The empirical research featured in this doctoral dissertation contributes to several scientific domains. The most important research contribution is in the domain of **Ethics & Morality**. The first contribution to this scientific domain consisted of the use of a unsupervised learning algorithm in a morally charged dataset (chapter 2). Learning algorithms are used for data analysis in many fields, but remain largely unexplored in the moral domain. This is an unfortunate missing opportunity, as these algorithms have the potential to further the knowledge of morality. The second contribution consisted of empirical information on particular ethical issues related to socially relevant AI Systems (chapters 3 and 4). A third contribution was made to the particular domain of artificial morality by reporting different perspectives about AMAs (chapter 5) and operationalizing part of the theoretical metanormative framework for decision making under moral uncertainty (chapter 6). This dissertation also contributes to the domain of **Discrete Choice Analysis**. By using a latent class choice model to operationalize the metanormative framework for decision making under moral uncertainty, the research featured in chapter 6 opens avenues for using choice analysis and models in artificial moral decision making. Finally, this dissertation also contributes to **Empirical Legal Studies**. Although not the primary focus of this research, using a unsupervised learning algorithm in a morally charged dataset of gender-based violence legal cases, led to significant findings about these legal decisions in cases decided in Portuguese courts (chapter 2).

7.2.3. CONTRIBUTIONS FOR FUTURE RESEARCH

Throughout this dissertation, several implications for future research were outlined. Rather than repeating those implications, which can be found in each chapter of the dissertation, **three key contributions** are presented: (i) three-dimensional construction of *Ethical AI*; (ii) value of empirical information in AI Ethics; and (iii) mitigation of speculation in the current data-driven paradigm in AI. These contributions are expected to be used in future AI Ethics research.

THE THREE-DIMENSIONAL CONSTRUCTION OF ETHICAL AI

AI is a disruptive technology entangled in normative issues. Theorizing about AI, Ethics, and Morality is particularly challenging due to the complexity and vastness of these topics. Rather than attempting at rigorous definitions, the focus of this dissertation was on the relationship between these topics. This relationship was conceptualized in **three different dimension**s: (i) AI as a data analysis tool in morality; (ii) AI as a system in a morally charged context; and (iii) AI as an artificial moral agent. Whereas other conceptualizations are possible, this three-dimensional construction of the relationship between AI and Ethics is practical yet comprehensive. It allows different socially relevant AI Systems in each

dimension encompassing a wide range of technologies, from relatively standard data analysis tools to sophisticated and even fictional artificial systems, thus promoting a rich normative debate. This construction may serve as a **reference tool** in further research endeavors in AI Ethics.

THE VALUE OF EMPIRICAL INFORMATION IN AI ETHICS

This doctoral dissertation was built on the premise that empirical information is valuable for AI Ethics to realize its normative mandate. As mentioned above, the research presented in this dissertation is well-suited to facilitate the reflective development of AI. This research illustrates the potential of using an **empirical approach in the morality domain without compromising the diversity of thought and normative edge**. A good example is the empirical study on AMAs, which systematically unravels contrasting views on the development, design, moral agency, and future projections of AMAs, thus allowing an accessible, structured, but also thought provocative debate on artificial morality. This dissertation **opens avenues for empirical research** in AI Ethics located along the spectrum of empirical ethics and morally relevant social sciences.

MITIGATING SPECULATION IN THE CURRENT DATA-DRIVEN PARADIGM IN AI

History has cautioned about unsubstantiated and speculative claims regarding AI. This doctoral dissertation was written in a period of great excitement about AI, and, to a great extent, it aimed at **grounding the normative debates about AI**. The AV is the illustrative case study of speculation in the current data-driven paradigm. Deep learning allowed significant technological advances in this field, but reckless claims about the deployment timeline fueled a speculative normative debate about autonomous driving. The ***AV trolley problem*** was reported back in the 1970s [115], but the projected advent of autonomous driving brought unprecedented attention to this thought experiment. The normative debates revolved around various aspects of the AV trolley problem without much consideration of the state-of-the-art technology or input from developers. The study featured in this doctoral dissertation on the ethical issues in focus by the AV industry aimed at contributing to a **meaningful normative debate about autonomous driving**. A wide range of ethical issues was conceptualized within the AV technology, and the industry's take on those issues, namely safety in extreme traffic situations, was presented. It is again emphasized that the AV trolley problem is an important thought experiment, not only because of the rich discussions reported in the literature but also because it raised widespread attention to the ethical issues associated with these technologies. However, the AV trolley problem is also yet another cautionary tale about speculation in AI, given its unrealistic premises which alienate other communities from engaging in the normative debate, and the fact that it steered attention from other pressing normative issues associated with autonomous driving. **Speculation weakens the normative endeavor of AI Ethics**. In the future, given that novel AI paradigms will likely emerge, and there is again a risk for speculation, these cautionary tales should be integrated early on in the normative debates.

7.3. POLICY REFLECTIONS

The exploratory nature of this research is not conducive to hard recommendations. Instead, a reflection on AI policy in light of the empirical work reported in the dissertation is more appropriate. For reasons of convenience, the focus of this reflection is limited to **European Union AI policy**.

The European Union's (EU) approach to AI is centered on the idea that it must be grounded in fundamental values and rights. There is a clear concern about establishing standards throughout the EU to ensure that AI is human-centric, ethical, and sustainable [461]. To realize this ambition, the **European Commission has developed several policy actions** in recent years. The actions are materialized in a series of policy documents that outline the strategic vision for the policy and regulatory framework on AI [1]. These preparatory documents are central in this reflection, while it is also acknowledged that additional policy and regulatory packages are expected in upcoming years.

The European Commission has decided on a policy entailing a regulatory framework for high-risk AI systems along with the possibility for all providers of non-risk AI systems to follow a code of conduct. By restricting regulatory requirements to systems that pose a high risk of violation of safety and fundamental rights, this policy leaves **plenty of room for normative guidance of these technologies and reinforces the need for a reflective development of AI**. The research featured in this dissertation seems to be well-aligned with the policy ambitions of the Commission as it makes an empirical contribution to such reflective development.

Across the current AI policy landscape, there is a clear emphasis on the **need to involve stakeholders**. The High Level Expert Group on AI (HLEG) cautioned the Commission to institutionalize a dialogue between policy-makers, developers, and users of technology (*Policy and Investment Recommendations for Trustworthy AI*). This group had also conducted several consultations with stakeholders when drafting the reports *Ethics Guidelines for Trustworthy AI* and *Assessment List for Trustworthy AI*. Accordingly, the Commission has carried actions in multi-stakeholder open platforms such as the European AI Alliance. For instance, upon publication of the white paper *On Artificial Intelligence - A European approach to excellence and trust*, the Commission sought structured feedback from stakeholders on the proposed policy options. These actions seem to go beyond standard public consultation in "right of initiative" procedures, and have received widespread participation.

These policy actions improve the communication with different stakeholders and are important to address the challenges related to the multitude of stakeholders operating in the AI space. Throughout this dissertation, it has been reiterated that there is a need for collaboration between the different AI stakeholders. It was posited that such collaboration should rely on clear communication of technology and normative insights. Through the actions mentioned above, the Commission has fostered a collaborative environment but **improvements are possible when it comes to fostering a reflective development of AI**.

Normative work is often communicated to developers and other stakeholders operating in the AI space as a top-down set of guidelines, checklists, or standards. An illustrative case is the *Assessment List for Trustworthy AI*, a practical checklist tool derived from the *Ethics Guidelines for Trustworthy AI and Policy*, which is designed to assist organizations to identify

[1] The European Commission has relied on expert reports such as *Ethics Guidelines for Trustworthy AI and Policy*; *Investment Recommendations for Trustworthy AI*; and *Ethics of Connected and Automated Vehicles*, to outline the policy vision, which was initially materialized in the *White Paper on Artificial Intelligence - A European approach to excellence and trust* and, subsequently, in a AI package consisting of a *Communication on Fostering a European Approach to Artificial Intelligence*, a *Coordinated Plan with Member States: 2021 update*, and a *Proposal for an AI Regulation*.

the risks associated with AI. This tool effectively translates principles into requirements, and it is instrumental for developers to comply with normative standards. However, it also reinforces a sort of "tick-the-box" AI Ethics and does little to assist stakeholders to engage in more nuanced normative critical thinking about AI. The HLEG acknowledges this limitation, as it cautions that *organizations derive the most value from the list by active engagement with the questions it raises, which aim at encouraging thoughtful reflection to provoke appropriate action and nurture an organizational culture committed to developing and maintaining Trustworthy AI systems.*

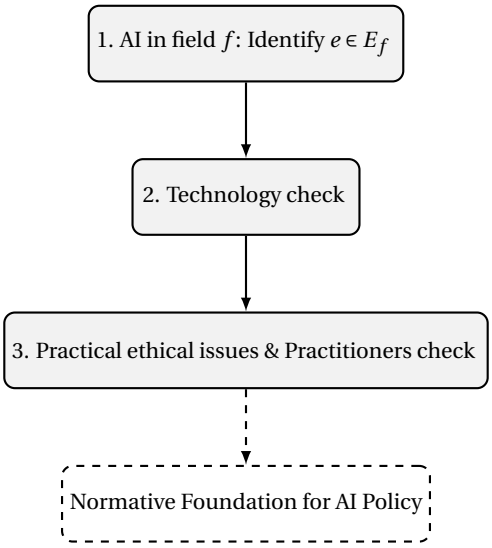
This dissertation made a case for an empirical approach in AI Ethics that is systematic without compromising the diversity of thought and normative edge. This approach is expected to assist stakeholders in engaging in a reflective development of AI. A practical **high level-approach** that may be used in **future AI policy actions** is presented (Figure 7.1). This three-step approach integrates theoretical and empirical insights from different communities and stakeholders operating in the AI space. It may be a stepping stone in building a rich normative foundation for future AI policies.

Step 1: Extreme Cases Upon the introduction of an AI-powered disruptive technology in a societal domain or field $f \in [F]$, normative discussions tend to focus on extreme cases $e \in [E]$. Ethics deals with values and moral principles. It often explores carefully crafted extreme moral dilemmas to test moral intuitions and further develop moral principles. In the AI domain, those dilemmas tend to feature an artificial super system set to take over difficult human moral decisions. The AV trolley problem is an illustrative e that has been widely debated. In this practical approach, extreme cases E are acknowledged and treated as normative boundaries. Within such boundaries, a normative contextualization and a grounded analysis of the technology in f should take place.

Step 2: Technology Check A technology check is required in order to ground the normative discussions of AI within boundaries E. The aim is to settle the state-of-the-art and realistic potential of the AI technology in f. This technology check may be realized through industry reviews, surveys, or interviews with developers.

Step 3: Practical Ethical Issues The myriad of guidelines issued by different AI stakeholders has led to the compilation of overarching ethical issues and principles related to AI technologies [151, 152]. These concepts are valuable tools for normative guidance, but they tend to be quite abstract and, therefore, need operationalization and further contextualization in f. Key issues, such as safety, accountability, human oversight, or explainability, have different implications in different societal domains. A theoretical normative reflection and insights of practitioners in f are relevant for such contextualization.

Figure 7.1: Practical approach for AI policy of systems operating in morally charged contexts.



An impressive amount of AI policy work has been developed in the EU in recent years. The European Commission and the HLEG have sought feedback from stakeholders, and their widespread participation signals great engagement in AI. Through these policy actions, the Commission has fostered a collaborative environment, but improvements are possible to foster a reflective development of AI. This three-step structured approach to AI policy incorporates practical insights from normative and technology communities while also acknowledging the diversity of thought and normative edge in AI Ethics, thus fostering a reflective development of AI. As AI policy is now moving into contextualizing AI in different societal domains, this approach may guide future AI policies.

7.4. FINAL REFLECTION FROM AUTHOR

This was a precious time to write a doctoral dissertation on AI. When I started my Ph.D. at TU Delft in early 2018, AI was about to burst in Europe. Later that year, a high level expert group was set up by the European Commission, which included TPM faculty. A couple of years later, the University launched several TU Delft AI Labs. The overall feeling about the potential of AI was refreshing, stimulating, and inspiring. Like many other researchers at TU Delft, I also felt compelled to work in AI.

My colleagues in the ERC-funded Behave research group were working on morality and choice behavior. Eventually, it became clear that I would focus on the moral aspects of AI while also taking advantage of the empirical culture of the group. I was convinced that in the current AI paradigm, empirical research was important for normative guidance.

For four years, I worked very hard on this ambitious project which aimed to provide empirical information on the different dimensions of *Ethical AI*. In the end, I believe I succeeded in making a small contribution to AI. As in every doctoral project, there were many challenges, trials, and tribulations. There was also a global pandemic that brought about so many

fears and changes in how we live and work. Nevertheless, looking back at my Ph.D. I see an immense joy in learning, many great social events, and a wonderful fellowship in my research group.

REFERENCES

[1] P. Gupta and P. Kumar, *The capacity of wireless networks,* IEEE Transactions on Information Theory **46**, 388 (2000).

[2] A. Goldsmith, *Wireless communications* (Cambridge university press, 2005).

[3] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning,* Nature **521**, 436 (2015).

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, 2016).

[5] G. E. Moore *et al.*, *Cramming more components onto integrated circuits,* (1965).

[6] C. E. Leiserson, N. C. Thompson, J. S. Emer, B. C. Kuszmaul, B. W. Lampson, D. Sanchez, and T. B. Schardl, *There's plenty of room at the top: What will drive computer performance after moore's law?* Science **368** (2020), 10.1126/science.aam9744, https://science.sciencemag.org/content/368/6495/eaam9744.full.pdf .

[7] M. Musib, F. Wang, M. A. Tarselli, R. Yoho, K.-H. Yu, R. M. Andrés, N. F. Greenwald, X. Pan, C.-H. Lee, J. Zhang, K. Dutton-Regester, J. W. Johnston, and I. M. Sharafeldin, *Artificial intelligence in research,* Science **357**, 28 (2017), https://science.sciencemag.org/content/357/6346/28.full.pdf .

[8] D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, and J. J. Collins, *Next-generation machine learning for biological networks,* Cell **173**, 1581 (2018).

[9] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, *Machine learning for molecular and materials science,* Nature **559**, 547 (2018).

[10] N. Artrith, K. T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain, and A. Walsh, *Best practices in machine learning for chemistry,* Nature Chemistry **13**, 505 (2021).

[11] E. Bedolla, L. C. Padierna, and R. Castañeda-Priego, *Machine learning for condensed matter physics,* Journal of Physics: Condensed Matter **33**, 053001 (2020).

[12] A. Radovic, M. Williams, D. Rousseau, M. Kagan, D. Bonacorsi, A. Himmel, A. Aurisano, K. Terao, and T. Wongjirad, *Machine learning at the energy and intensity frontiers of particle physics,* Nature **560**, 41 (2018).

[13] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, *Highly accurate protein structure prediction with alphafold,* Nature **596**, 583 (2021).

[14] R. Abduljabbar, H. Dia, S. Liyanage, and S. A. Bagloee, *Applications of artificial intelligence in transport: An overview,* Sustainability **11** (2019), 10.3390/su11010189.

[15] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixão, F. Mutz, L. de Paula Veronese, T. Oliveira-Santos, and A. F. De Souza, *Self-driving cars: A survey,* Expert Systems with Applications **165**, 113816 (2021).

[16] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, *A review of deep learning for renewable energy forecasting,* Energy Conversion and Management **198**, 111799 (2019).

[17] M. Bourdeau, X. qiang Zhai, E. Nefzaoui, X. Guo, and P. Chatellier, *Modeling and forecasting building energy consumption: A review of data-driven techniques,* Sustainable Cities and Society **48**, 101533 (2019).

[18] Y. Zhao, T. Li, X. Zhang, and C. Zhang, *Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future,* Renewable and Sustainable Energy Reviews **109**, 85 (2019).

[19] K. A. Severson, P. M. Attia, N. Jin, N. Perkins, B. Jiang, Z. Yang, M. H. Chen, M. Aykol, P. K. Herring, D. Fraggedakis, M. Z. Bazant, S. J. Harris, W. C. Chueh, and R. D. Braatz, *Data-driven prediction of battery cycle life before capacity degradation,* Nature Energy **4**, 383 (2019).

[20] M. Taddeo, T. McCutcheon, and L. Floridi, *Trusting artificial intelligence in cybersecurity is a double-edged sword,* Nature Machine Intelligence **1**, 557 (2019).

[21] J.-h. Li, *Cyber security meets artificial intelligence: a survey,* Frontiers of Information Technology & Electronic Engineering **19**, 1462 (2018).

[22] Z. Sabetsarvestani, B. Sober, C. Higgitt, I. Daubechies, and M. R. D. Rodrigues, *Artificial intelligence for art investigation: Meeting the challenge of separating x-ray images of the ghent altarpiece,* Science Advances **5** (2019), 10.1126/sciadv.aaw7416, https://advances.sciencemag.org/content/5/8/eaaw7416.full.pdf .

[23] M. Mazzone and A. Elgammal, *Art, creativity, and the potential of artificial intelligence,* Arts **8** (2019), 10.3390/arts8010026.

[24] B. L. T. Sturm, M. Iglesias, O. Ben-Tal, M. Miron, and E. Gómez, *Artificial intelligence and music: Open questions of copyright law and engineering praxis,* Arts **8** (2019), 10.3390/arts8030115.

[25] S. Audry and J. Ippolito, *Can artificial intelligence make art without artists? ask the viewer,* Arts **8** (2019), 10.3390/arts8010035.

[26] R. Arkin, *Governing lethal behavior in autonomous robots* (Chapman and Hall/CRC, 2009).

[27] G. M. Campedelli, M. Bartulovic, and K. M. Carley, *Learning future terrorist targets through temporal meta-graphs,* Scientific Reports **11**, 8533 (2021).

[28] B. Ganor, *Artificial or human: A new era of counterterrorism intelligence?* Studies in Conflict & Terrorism **44**, 605 (2021), https://doi.org/10.1080/1057610X.2019.1568815 .

[29] B. K. Wiederhold, *Internet dating: Should you try it?* Cyberpsychology, Behavior, and Social Networking **23**, 195 (2020), pMID: 32271129, https://doi.org/10.1089/cyber.2020.29178.bkw .

[30] V. Larivière, S. Haustein, and P. Mongeon, *The oligopoly of academic publishers in the digital era,* PloS one **10**, e0127502 (2015).

[31] P. Langley, *The changing science of machine learning,* (2011).

[32] J. Haugeland, *Artificial intelligence: The very idea* (MIT press, 1989).

[33] V. Dignum, *Responsible artificial intelligence: how to develop and use AI in a responsible way* (Springer Nature, 2019).

[34] S. M. Liao, *Ethics of Artificial Intelligence* (Oxford University Press, 2020).

[35] B. Logan, *A future for agent programming,* in *Engineering Multi-Agent Systems*, edited by M. Baldoni, L. Baresi, and M. Dastani (Springer International Publishing, Cham, 2015) pp. 3–17.

[36] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, *Formal verification of ethical choices in autonomous systems,* Robotics and Autonomous Systems **77**, 1 (2016).

[37] M. Garnelo and M. Shanahan, *Reconciling deep learning with symbolic artificial intelligence: representing objects and relations,* Current Opinion in Behavioral Sciences **29**, 17 (2019), artificial Intelligence.

[38] V. Dignum and F. Dignum, *Agents are dead. long live agents!* in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (2020) pp. 1701–1705.

[39] C. Baroglio, J. F. Hubner, and M. Winikoff, *Engineering multi-agent systems,* in *Proceedings of the 8th International Workshop on Engineering Multi-Agent Systems* (Springer, 2020).

[40] L. A. Dennis and N. Oren, *Explaining bdi agent behaviour through dialogue,* in *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)* (International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2021).

[41] J. Dix, B. Logan, and M. Winikoff, *Preface to the special issue on engineering reliable multi-agent systems,* Autonomous Agents and Multi-Agent Systems **35**, 37 (2021).

[42] T. M. Mitchell, *Machine learning*, edited by E. Munson (McGraw-hill New York, 1997).

[43] C. Cortes and V. Vapnik, *Support-vector networks,* Machine Learning **20**, 273 (1995).

[44] I. Rish *et al.*, *An empirical study of the naive bayes classifier,* in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3 (2001) pp. 41–46.

[45] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities,* Proceedings of the National Academy of Sciences of the United States of America **79**, 2554 (1982).

[46] Y. LeCun, Y. Bengio, *et al.*, *Convolutional networks for images, speech, and time series,* The handbook of brain theory and neural networks **3361**, 1995 (1995).

[47] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning representations by back-propagating errors,* Nature **323**, 533 (1986).

[48] K. de Vries, *Identity, profiling algorithms and a world of ambient intelligence,* Ethics and Information Technology **12**, 71 (2010).

[49] C.-K. Kao and D. M. Liebovitz, *Consumer mobile health apps: Current state, barriers, and future directions,* PM&R **9**, S106 (2017), clinical Informatics in Physiatry.

[50] J. K. Carroll, A. Moorhead, R. Bond, W. G. LeBlanc, R. J. Petrella, and K. Fiscella, *Who uses mobile phone health apps and does use matter? a secondary data analytics approach,* J Med Internet Res **19**, e125 (2017).

[51] W. J. Gordon, A. Landman, H. Zhang, and D. W. Bates, *Beyond validation: getting health apps into clinical practice,* npj Digital Medicine **3**, 14 (2020).

[52] B. Babic, S. Gerke, T. Evgeniou, and I. G. Cohen, *Direct-to-consumer medical machine learning and artificial intelligence applications,* Nature Machine Intelligence **3**, 283 (2021).

[53] J. McGill, S. Bouclin, and A. Salyzyn, *Mobile and web-based legal apps: Opportunities, risks and information gaps,* Risks and Information Gaps (April 28, 2017) **15** (2017).

[54] T. Sourdin, J. Meredith, and B. Li, *Digital technology and justice: justice apps* (Routledge, 2020).

[55] E. S. T. Poppe, *The future is complicated: Ai, apps & access to justice,* Okla. L. Rev. **72**, 185 (2019).

[56] P. B. Brandtzaeg and A. Følstad, *Why people use chatbots,* in *Internet Science*, edited by I. Kompatsiaris, J. Cave, A. Satsiou, G. Carle, A. Passani, E. Kontopoulos, S. Diplaris, and D. McMillan (Springer International Publishing, Cham, 2017) pp. 377–392.

[57] E. Adamopoulou and L. Moussiades, *An overview of chatbot technology,* in *Artificial Intelligence Applications and Innovations*, edited by I. Maglogiannis, L. Iliadis, and E. Pimenidis (Springer International Publishing, Cham, 2020) pp. 373–383.

[58] A. Følstad, C. B. Nordheim, and C. A. Bjørkli, *What makes users trust a chatbot for customer service? an exploratory interview study,* in *Internet Science*, edited by S. S. Bodrunova (Springer International Publishing, Cham, 2018) pp. 194–208.

[59] M. Chung, E. Ko, H. Joung, and S. J. Kim, *Chatbot e-service and customer satisfaction regarding luxury brands,* Journal of Business Research **117**, 587 (2020).

[60] T. Makasi, A. Nili, K. C. Desouza, and M. Tate, *A typology of chatbots in public service delivery,* IEEE Software , 0 (2021).

[61] A. Androutsopoulou, N. Karacapilidis, E. Loukis, and Y. Charalabidis, *Transforming the communication between citizens and government through ai-guided chatbots,* Government Information Quarterly **36**, 358 (2019).

[62] N. Aoki, *An experimental study of public trust in ai chatbots in the public sector,* Government Information Quarterly **37**, 101490 (2020).

[63] A. Følstad, M. Skjuve, and P. B. Brandtzaeg, *Different chatbots for different purposes: Towards a typology of chatbots to understand interaction design,* in *Internet Science,* edited by S. S. Bodrunova, O. Koltsova, A. Følstad, H. Halpin, P. Kolozaridi, L. Yuldashev, A. Smoliarova, and H. Niedermayer (Springer International Publishing, Cham, 2019) pp. 145–156.

[64] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, *Chatbots and conversational agents in mental health: A review of the psychiatric landscape,* The Canadian Journal of Psychiatry **64**, 456 (2019), pMID: 30897957, https://doi.org/10.1177/0706743719828977 .

[65] A. Palanica, P. Flaschner, A. Thommandram, M. Li, and Y. Fossat, *Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey,* J Med Internet Res **21**, e12887 (2019).

[66] J. Wirtz, P. G. Patterson, W. H. Kunz, T. Gruber, V. N. Lu, S. Paluch, and A. Martins, *Brave new world: service robots in the frontline,* Journal of Service Management **29**, 907 (2021).

[67] J. Forlizzi and C. DiSalvo, *Service robots in the domestic environment: A study of the roomba vacuum in the home,* in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, HRI '06 (Association for Computing Machinery, New York, NY, USA, 2006) p. 258–265.

[68] H. Robinson, B. MacDonald, and E. Broadbent, *The role of healthcare robots for older people at home: A review,* International Journal of Social Robotics **6**, 575 (2014).

[69] S. Frennert, H. Aminoff, and B. ostlund, *Technological frames and care robots in eldercare,* International Journal of Social Robotics **13**, 311 (2021).

[70] W. Brenner and A. Herrmann, *An overview of technology, benefits and impact of automated and autonomous driving on the automotive industry,* in *Digital Marketplaces Unleashed*, edited by C. Linnhoff-Popien, R. Schneider, and M. Zaddach (Springer Berlin Heidelberg, Berlin, Heidelberg, 2018) pp. 427–442.

[71] D. P. Piatkowski, *Autonomous shuttles: What do users expect and how will they use them?* Journal of Urban Technology **0**, 1 (2021), https://doi.org/10.1080/10630732.2021.1896345 .

[72] J. Fleetwood, *Public health, ethics, and autonomous vehicles,* American journal of public health **107**, 532 (2017).

[73] W. Schwarting, J. Alonso-Mora, and D. Rus, *Planning and decision-making for autonomous vehicles,* Annual Review of Control, Robotics, and Autonomous Systems **1**, 187 (2018), https://doi.org/10.1146/annurev-control-060117-105157 .

[74] J. Janai, F. Güney, A. Behl, and A. Geiger, *Computer vision for autonomous vehicles: Problems, datasets and state of the art,* Foundations and Trends® in Computer Graphics and Vision **12**, 1 (2020).

[75] A. Qayyum, M. Usama, J. Qadir, and A. Al-Fuqaha, *Securing connected amp; autonomous vehicles: Challenges posed by adversarial machine learning and the way forward,* IEEE Communications Surveys Tutorials **22**, 998 (2020).

[76] D. Milakis, B. Van Arem, and B. Van Wee, *Policy and society related implications of automated driving: A review of literature and directions for future research,* Journal of Intelligent Transportation Systems **21**, 324 (2017).

[77] R. Vinuesa, H. Azizpour, I. Leite, M. Balaam, V. Dignum, S. Domisch, A. Felländer, S. D. Langhans, M. Tegmark, and F. Fuso Nerini, *The role of artificial intelligence in achieving the sustainable development goals,* Nature Communications **11**, 233 (2020).

[78] J. Maclure and S. Russell, *Ai for humanity: The global challenges,* in *Reflections on Artificial Intelligence for Humanity*, edited by B. Braunschweig and M. Ghallab (Springer International Publishing, Cham, 2021) pp. 116–126.

[79] J. Guo and B. Li, *The application of medical artificial intelligence technology in rural areas of developing countries,* Health Equity **2**, 174 (2018), pMID: 30283865, https://doi.org/10.1089/heq.2018.0037 .

[80] B. Wahl, A. Cossy-Gantner, S. Germann, and N. R. Schwalbe, *Artificial intelligence (ai) and global health: how can ai contribute to health in resource-poor settings?* BMJ Global Health **3** (2018), 10.1136/bmjgh-2018-000798, https://gh.bmj.com/content/3/4/e000798.full.pdf .

[81] N. Schwalbe and B. Wahl, *Artificial intelligence and the future of global health,* The Lancet **395**, 1579 (2020).

[82] K. Fuhad, J. F. Tuba, M. Sarker, R. Ali, S. Momen, N. Mohammed, and T. Rahman, *Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application,* Diagnostics **10**, 329 (2020).

[83] J. West and M. Bhattacharya, *Intelligent financial fraud detection: a comprehensive review,* Computers & security **57**, 47 (2016).

[84] A. Meijer and M. Wessels, *Predictive policing: Review of benefits and drawbacks,* International Journal of Public Administration **42**, 1031 (2019).

[85] P. Hajek and R. Henriques, *Mining corporate annual reports for intelligent detection of financial statement fraud–a comparative study of machine learning methods,* Knowledge-Based Systems **128**, 139 (2017).

[86] A. W. Flores, K. Bechtel, and C. T. Lowenkamp, *False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks,* Fed. Probation **80**, 38 (2016).

[87] D. Leslie, *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of ai systems in the public sector,* Available at SSRN 3403301 (2019).

[88] R. B. Parikh, S. Teeple, and A. S. Navathe, *Addressing bias in artificial intelligence in health care,* Jama **322**, 2377 (2019).

[89] G. S. Nelson, *Bias in artificial intelligence,* North Carolina Medical Journal **80**, 220 (2019), https://www.ncmedicaljournal.com/content/80/4/220.full.pdf .

[90] M. DeCamp and C. Lindvall, *Latent bias and the implementation of artificial intelligence in medicine,* Journal of the American Medical Informatics Association **27**, 2020 (2020).

[91] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, *From what to how: An initial review of publicly available ai ethics tools, methods and research to translate principles into practices,* Science and Engineering Ethics **26**, 2141 (2020).

[92] A. Taeihagh and H. S. M. Lim, *Towards autonomous vehicles in smart cities: Risks and risk governance,* in *Towards Connected and Autonomous Vehicle Highways: Technical, Security and Social Challenges*, edited by U. Z. A. Hamid and F. Al-Turjman (Springer International Publishing, Cham, 2021) pp. 169–190.

[93] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, *A survey on bias and fairness in machine learning,* ACM Comput. Surv. **54** (2021), 10.1145/3457607.

[94] B. Schoettle and M. Sivak, *A preliminary analysis of real-world crashes involving self-driving vehicles,* University of Michigan Transportation Research Institute (2015).

[95] F. Favarò, S. Eurich, and N. Nader, *Autonomous vehicles' disengagements: Trends, triggers, and regulatory limitations,* Accident Analysis & Prevention **110**, 136 (2018).

[96] W. Biever, L. Angell, and S. Seaman, *Automated driving system collisions: Early lessons,* Human Factors **62**, 249 (2020), pMID: 31502899, https://doi.org/10.1177/0018720819872034 .

[97] F. M. Favarò, N. Nader, S. O. Eurich, M. Tripp, and N. Varadaraju, *Examining accident reports involving autonomous vehicles in california,* PLoS one **12**, e0184952 (2017).

[98] S. Nyholm, *The ethics of crashes with self-driving cars: A roadmap, i,* Philosophy Compass **13**, e12507 (2018), e12507 PHCO-1155, https://onlinelibrary.wiley.com/doi/pdf/10.1111/phc3.12507 .

[99] J. Claybrook and S. Kildare, *Autonomous vehicles: No driver... no regulation?* Science **361**, 36 (2018), https://science.sciencemag.org/content/361/6397/36.full.pdf .

[100] A. Maedche, C. Legner, A. Benlian, B. Berger, H. Gimpel, T. Hess, O. Hinz, S. Morana, and M. Söllner, *Ai-based digital assistants,* Business & Information Systems Engineering , 1 (2019).

[101] J. Dastin, *Amazon scraps secret ai recruiting tool that showed bias against women. reuters, october 2018,* (2018).

[102] L. N. Guo, M. S. Lee, B. Kassamali, C. Mita, and V. E. Nambudiri, *Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection—a scoping review,* Journal of the American Academy of Dermatology (2021), https://doi.org/10.1016/j.jaad.2021.06.884.

[103] C. Rudin, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,* Nature Machine Intelligence **1**, 206 (2019).

[104] X. Dastile, T. Celik, and M. Potsane, *Statistical and machine learning models in credit scoring: A systematic literature survey,* Applied Soft Computing **91**, 106263 (2020).

[105] M. Bücker, G. Szepannek, A. Gosiewska, and P. Biecek, *Transparency, auditability, and explainability of machine learning models in credit scoring,* Journal of the Operational Research Society **0**, 1 (2021), https://doi.org/10.1080/01605682.2021.1922098 .

[106] K. Warwick, *Artificial intelligence: the basics*, edited by Routledge (Routledge, 2013).

[107] A. M. Turing and J. Haugeland, *Computing machinery and intelligence* (MIT Press Cambridge, MA, 1950).

[108] N. Wiener, *God and Golem, Inc: A Comment on Certain Points where Cybernetics Impinges on Religion*, Vol. 42 (MIT press, 1964).

[109] N. Wiener, *Some moral and technical consequences of automation,* Science **131**, 1355 (1960).

[110] A. L. Samuel, *Some moral and technical consequences of automation—a refutation,* Science **132**, 741 (1960).

[111] H. L. Dreyfus, *What computers can't do : a critique of artificial reason* (Harper & Row, 1972).

[112] H. L. Dreyfus, L. Hubert, *et al.*, *What computers still can't do: A critique of artificial reason* (MIT press, 1992).

[113] J. R. Searle, *Minds, brains, and programs,* Behavioral and Brain Sciences **3**, 417–424 (1980).

[114] J. McCarthy, *Generality in artificial intelligence,* Commun. ACM **30**, 1030–1035 (1987).

[115] O. Firschein, M. A. Fischler, L. S. Coles, and J. M. Tenenbaum, *Forecasting and assessing the impact of artificial intelligence on society,* in *IJCAI*, Vol. 5 (Citeseer, 1973) pp. 105–120.

[116] E. A. Feigenbaum, *Knowledge Engineering: The Applied Side of Artificial Intelligence.*, Tech. Rep. (STANFORD UNIV CA DEPT OF COMPUTER SCIENCE, 1980).

[117] G. Brown and K. Kornmayer, *Expert systems restructure managed care practice: implementation and ethics,* Behavioral healthcare tomorrow **5**, 31—34 (1996).

[118] K. Cass, *Expert systems as general-use advisory tools: An examination of moral responsibility,* Business and Professional Ethics Journal **15**, 61 (1996).

[119] J. F. Dillard and K. Yuthas, *A responsibility-based approach to the development of expert systems,* in *Proceedings of the International Association for Business and Society*, Vol. 7 (1996) pp. 351–360.

[120] Shu-Hsien Liao, *Expert system methodologies and applications—a decade review from 1995 to 2004,* Expert Systems with Applications **28**, 93 (2005).

[121] H. Heathfield, *The rise and 'fall' of expert systems in medicine,* Expert Systems **16**, 183 (1999), https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0394.00107 .

[122] P. Leith, *The rise and fall of the legal expert system,* International Review of Law, Computers & Technology **30**, 94 (2016), https://doi.org/10.1080/13600869.2016.1232465 .

[123] H. Moravec, *Mind children: The future of robot and human intelligence* (Harvard University Press, 1988).

[124] V. Vinge, *Technological singularity,* in *VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute* (1993) pp. 30–31.

[125] N. Bostrom, *Ethical issues in advanced artificial intelligence,* Science fiction and philosophy: from time travel to superintelligence , 277 (2003).

[126] E. Yudkowsky, *Creating friendly ai 1.0: The analysis and design of benevolent goal architectures,* The Singularity Institute, San Francisco, USA (2001).

[127] M. Anderson, S. L. Anderson, and C. Armen, *Towards machine ethics,* in *AAAI-04 workshop on agent organizations: theory and practice, San Jose, CA* (2004).

[128] J. H. Moor, *The nature, importance, and difficulty of machine ethics,* IEEE intelligent systems **21**, 18 (2006).

[129] C. Allen, I. Smit, and W. Wallach, *Artificial morality: Top-down, bottom-up, and hybrid approaches,* Ethics and information technology **7**, 149 (2005).

[130] M. Anderson and S. L. Anderson, *The status of machine ethics: a report from the aaai symposium,* Minds and Machines **17**, 1 (2007).

[131] M. Anderson and S. L. Anderson, *Machine ethics: Creating an ethical intelligent agent,* in *Machine Ethics and Robot Ethics* (Routledge, 2020) pp. 237–248.

[132] C. Allen, W. Wallach, and I. Smit, *Why machine ethics?* IEEE Intelligent Systems **21**, 12 (2006).

[133] J.-F. Bonnefon, A. Shariff, and I. Rahwan, *The social dilemma of autonomous vehicles,* Science **352**, 1573 (2016).

[134] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, *The moral machine experiment,* Nature **563**, 59 (2018).

[135] N. J. Goodall, *Machine ethics and automated vehicles,* in *Road Vehicle Automation*, edited by G. Meyer and S. Beiker (Springer International Publishing, Cham, 2014) pp. 93–102.

[136] L. Greenemeier, *Driverless cars will face moral dilemmas,* Scientific American (2016).

[137] F. S. de Sio, *Killing by autonomous vehicles and the legal doctrine of necessity,* Ethical Theory and Moral Practice **20**, 411 (2017).

[138] R. Sparrow and M. Howard, *When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport,* Transportation Research Part C: Emerging Technologies **80**, 206 (2017).

[139] N. JafariNaimi, *Our bodies in the trolley's path, or why self-driving cars must *not* be programmed to kill,* Science, Technology, & Human Values **43**, 302 (2018), https://doi.org/10.1177/0162243917718942 .

[140] J. Himmelreich, *Never mind the trolley: The ethics of autonomous vehicles in mundane situations,* Ethical Theory and Moral Practice , 1 (2018).

[141] G. Keeling, *Why trolley problems matter for the ethics of automated vehicles,* Science and engineering ethics , 1 (2019).

[142] A. Wolkenstein, *What has the trolley dilemma ever done for us (and what will it do in the future)? on some recent debates about the ethics of self-driving cars,* Ethics and Information Technology , 1 (2018).

[143] J. C. Gerdes and S. M. Thornton, *Implementable ethics for autonomous vehicles,* in *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*, edited by M. Maurer, J. C. Gerdes, B. Lenz, and H. Winner (Springer Berlin Heidelberg, Berlin, Heidelberg, 2015) pp. 87–102.

[144] P. Lin, *Why ethics matters for autonomous cars,* in *Autonomous driving* (Springer, Berlin, Heidelberg, 2016) pp. 69–85.

[145] D. Leben, *A rawlsian algorithm for autonomous vehicles,* Ethics and Information Technology **19**, 107 (2017).

[146] S. M. Thornton, S. Pan, S. M. Erlien, and J. C. Gerdes, *Incorporating ethical considerations into automated vehicle control,* IEEE Transactions on Intelligent Transportation Systems **18**, 1429 (2017).

[147] L. T. Bergmann, L. Schlicht, C. Meixner, P. König, G. Pipa, S. Boshammer, and A. Stephan, *Autonomous vehicles require socio-political acceptance—an empirical and philosophical perspective on the problem of moral decision making,* Frontiers in behavioral neuroscience **12**, 31 (2018).

[148] M. Cunneen, M. Mullins, F. Murphy, D. Shannon, I. Furxhi, and C. Ryan, *Autonomous vehicles and avoiding the trolley (dilemma): Vehicle perception, classification, and the challenges of framing decision ethics,* Cybernetics and Systems **51**, 59 (2020), https://doi.org/10.1080/01969722.2019.1660541 .

[149] E. Awad, S. Dsouza, A. Shariff, I. Rahwan, and J.-F. Bonnefon, *Universals and variations in moral decisions made in 42 countries by 70,000 participants,* Proceedings of the National Academy of Sciences **117**, 2332 (2020).

[150] Y. E. Bigman and K. Gray, *Life and death decisions of autonomous vehicles,* Nature **579**, E1 (2020).

[151] A. Jobin, M. Ienca, and E. Vayena, *The global landscape of ai ethics guidelines,* Nature Machine Intelligence **1**, 389 (2019).

[152] T. Hagendorff, *The ethics of ai ethics: An evaluation of guidelines,* Minds and Machines , 1 (2020).

[153] L. Floridi and J. Cowls, *A unified framework of five principles for ai in society,* Available at SSRN 3831321 (2019).

[154] J. Mökander, J. Morley, M. Taddeo, and L. Floridi, *Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations,* Science and Engineering Ethics **27**, 44 (2021).

[155] A. Jobin, M. Ienca, and E. Vayena, *The global landscape of ai ethics guidelines,* Nature Machine Intelligence **1**, 389 (2019).

[156] B. Mittelstadt, *Principles alone cannot guarantee ethical ai,* Nature Machine Intelligence **1**, 501 (2019).

[157] D. Schiff, J. Biddle, J. Borenstein, and K. Laas, *What's next for ai ethics, policy, and governance? a global overview,* in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20 (Association for Computing Machinery, New York, NY, USA, 2020) p. 153–158.

[158] J. Morley, A. Elhalal, F. Garcia, L. Kinsey, J. Mökander, and L. Floridi, *Ethics as a service: A pragmatic operationalisation of ai ethics,* Minds and Machines **31**, 239 (2021).

[159] L. Floridi, *Translating principles into practices of digital ethics: Five risks of being unethical,* Philosophy & Technology **32**, 185 (2019).

[160] M. Anderson and S. Anderson, *Machine Ethics* (Cambridge University Press, 2011).

[161] A. Etzioni and O. Etzioni, *Incorporating ethics into artificial intelligence,* The Journal of Ethics **21**, 403 (2017).

[162] J.-A. Cervantes, S. López, L.-F. Rodríguez, S. Cervantes, F. Cervantes, and F. Ramos, *Artificial moral agents: A survey of the current status,* Science and Engineering Ethics **26**, 501 (2020).

[163] J.-F. Bonnefon, A. Shariff, and I. Rahwan, *The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view],* Proceedings of the IEEE **107**, 502 (2019).

[164] M. Whittaker, K. Crawford, R. Dobbe, G. Fried, E. Kaziunas, V. Mathur, S. M. West, R. Richardson, J. Schultz, and O. Schwartz, *AI now report 2018* (AI Now Institute at New York University New York, 2018).

[165] A. McNamara, J. Smith, and E. Murphy-Hill, *Does acm's code of ethics change ethical decision making in software development?* in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2018 (Association for Computing Machinery, New York, NY, USA, 2018) p. 729–733.

[166] W. Orr and J. L. Davis, *Attributions of ethical responsibility by artificial intelligence practitioners,* Information, Communication & Society **23**, 719 (2020), https://doi.org/10.1080/1369118X.2020.1713842 .

[167] A. Martinho, N. Herber, M. Kroesen, and C. Chorus, *Ethical issues in focus by the autonomous vehicles industry,* Transport reviews **41**, 556 (2021).

[168] P. Terzis, *Onward for the freedom of others: Marching beyond the ai ethics,* in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20 (Association for Computing Machinery, New York, NY, USA, 2020) p. 220–229.

[169] S. Fazelpour and Z. C. Lipton, *Algorithmic fairness from a non-ideal perspective,* in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20 (Association for Computing Machinery, New York, NY, USA, 2020) p. 57–63.

[170] P. BORRY, P. SCHOTSMANS, and K. DIERICKX, *The birth of the empirical turn in bioethics,* Bioethics **19**, 49 (2005), https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8519.2005.00424.x .

[171] R. De Vries and B. Gordijn, *Empirical ethics and its alleged meta-ethical fallacies,* Bioethics **23**, 193 (2009), https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8519.2009.01710.x .

[172] B. Molewijk, A. M. Stiggelbout, W. Otten, H. M. Dupuis, and J. Kievit, *Scientific contribution. empirical data and moral theory. a plea for integrated empirical ethics,* Medicine, Health Care and Philosophy **7**, 55 (2004).

[173] A. W. Musschenga, *Empirical Ethics, Context-Sensitivity, and Contextualism,* The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine **30**, 467 (2005), https://academic.oup.com/jmp/article-pdf/30/5/467/2729155/30-5-467.pdf .

[174] M. PARKER, *Two concepts of empirical ethics,* Bioethics **23**, 202 (2009), https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8519.2009.01708.x .

[175] M. Dunn, M. Sheehan, T. Hope, and M. Parker, *Toward methodological innovation in empirical ethics research,* Cambridge Q. Healthcare Ethics **21**, 466 (2012).

[176] G. R. Weaver and L. K. Trevino, *Normative and empirical business ethics: Separation, marriage of convenience, or marriage of necessity?* Business Ethics Quarterly , 129 (1994).

[177] L. van der Scheer and G. Widdershoven, *Integrated empirical ethics: Loss of normativity?* Medicine, Health Care and Philosophy **7**, 71 (2004).

[178] E. J. Topol, *High-performance medicine: the convergence of human and artificial intelligence,* Nature medicine **25**, 44 (2019).

[179] M.-C. Laï, M. Brian, and M.-F. Mamzer, *Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in france,* Journal of Translational Medicine **18**, 14 (2020).

[180] C. Blease, T. J. Kaptchuk, M. H. Bernstein, K. D. Mandl, J. D. Halamka, and C. M. DesRoches, *Artificial intelligence and the future of primary care: Exploratory qualitative study of uk general practitioners' views,* J Med Internet Res **21**, e12802 (2019).

[181] S. Oh, J. H. Kim, S.-W. Choi, H. J. Lee, J. Hong, and S. H. Kwon, *Physician confidence in artificial intelligence: an online mobile survey,* Journal of medical Internet research **21**, e12422 (2019).

[182] W. Wallach and C. Allen, *Moral machines: Teaching robots right from wrong* (Oxford University Press, 2008).

[183] A. van Wynsberghe and S. Robbins, *Critiquing the reasons for making artificial moral agents,* Science and Engineering Ethics **25**, 719 (2019).

[184] C. G. Chorus, B. Pudāne, N. Mouter, and D. Campbell, *Taboo trade-off aversion: A discrete choice model and empirical analysis,* Journal of choice modelling **27**, 37 (2018).

[185] S. Street, *A darwinian dilemma for realist theories of value,* Philosophical studies **127**, 109 (2006).

[186] J. D. Greene, *The secret joke of kant's soul,* Moral psychology **3**, 35 (2008).

[187] K. de Lazari-Radek and P. Singer, *The objectivity of ethics and the unity of practical reason,* Ethics **123**, 9 (2012).

[188] R. Joyce, *The evolution of morality* (MIT press, 2007).

[189] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto, *Moral foundations theory: The pragmatic validity of moral pluralism,* in *Advances in experimental social psychology*, Vol. 47 (Elsevier, 2013) pp. 55–130.

[190] D. S. Smith, *When two tribes go to law: The moral foundations theory and the brexit negotiations,* Forensic Science International: Mind and Law **2**, 100055 (2021).

[191] J. Graham, B. A. Nosek, J. Haidt, R. Iyer, S. Koleva, and P. H. Ditto, *Mapping the moral domain.* Journal of personality and social psychology **101**, 366 (2011).

[192] J. Graham, J. Haidt, and B. A. Nosek, *Liberals and conservatives rely on different sets of moral foundations.* Journal of personality and social psychology **96**, 1029 (2009).

[193] J. Haidt, C. Joseph, *et al.*, *The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules,* The innate mind **3**, 367 (2007).

[194] J. Haidt and C. Joseph, *Intuitive ethics: How innately prepared intuitions generate culturally variable virtues,* Daedalus **133**, 55 (2004).

[195] J. Haidt and J. Graham, *When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize,* Social Justice Research **20**, 98 (2007).

[196] J. Graham, B. A. Nosek, and J. Haidt, *The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum,* PLOS ONE **7**, 1 (2012).

[197] S. Clifford and J. Jerit, *How words do the work of politics: Moral foundations theory and the debate over stem cell research,* The Journal of Politics **75**, 659 (2013), https://doi.org/10.1017/S0022381613000492 .

[198] S. P. Koleva, J. Graham, R. Iyer, P. H. Ditto, and J. Haidt, *Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes,* Journal of research in personality **46**, 184 (2012).

[199] T. Grover, E. Bayraktaroglu, G. Mark, and E. H. R. Rho, *Moral and affective differences in u.s. immigration policy debate on twitter,* Computer Supported Cooperative Work (CSCW) **28**, 317 (2019).

[200] T. O'Grady, D. Vandegrift, M. Wolek, and G. Burr, *On the determinants of other-regarding behavior: Field tests of the moral foundations questionnaire,* Journal of Research in Personality **81**, 224 (2019).

[201] A. Nilsson, A. Erlandsson, and D. Västfjäll, *The congruency between moral foundations and intentions to donate, self-reported donations, and actual donations to charity,* Journal of Research in Personality **65**, 22 (2016).

[202] T. G. van den Berg, M. Kroesen, and C. G. Chorus, *Does morality predict aggressive driving? a conceptual analysis and exploratory empirical investigation,* Transportation Research Part F: Traffic Psychology and Behaviour **74**, 259 (2020).

[203] T. J. Vaughan, L. Bell Holleran, and J. R. Silver, *Applying moral foundations theory to the explanation of capital jurors' sentencing decisions,* Justice Quarterly **36**, 1176 (2019).

[204] T. L. L. Milfont, C. L. Davies, and M. S. Wilson, *The moral foundations of environmentalism: Care- and fairness-based morality interact with political liberalism to predict pro-environmental actions,* Social Psychological Bulletin **14**, 1 (2019).

[205] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, G. Moreno, C. Park, T. E. Chang, J. Chin, C. Leong, J. Y. Leung, A. Mirinjian, and M. Dehghani, *Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment,* Social Psychological and Personality Science **11**, 1057 (2020), https://doi.org/10.1177/1948550619876629 .

[206] J. Haidt, *The righteous mind: Why good people are divided by politics and religion* (Vintage, 2012).

[207] C. A. Harper and A. J. Harris, *Applying moral foundations theory to understanding public views of sexual offending,* Journal of Sexual Aggression **23**, 111 (2017), https://doi.org/10.1080/13552600.2016.1217086 .

[208] A. R. Pereira, D. N. Vieira, and T. Magalhães, *Fatal intimate partner violence against women in portugal: A forensic medical national study,* Journal of Forensic and Legal Medicine **20**, 1099 (2013).

[209] J. Quintas, P. Sousa, and G. Carolina, *As Respostas Judiciais na Criminalidade de Género* (Almedina, 2021).

[210] C. Machado, M. Gonçalves, M. Matos, and A. R. Dias, *Child and partner abuse: Self-reported prevalence and attitudes in the north of portugal,* Child Abuse & Neglect **31**, 657 (2007).

[211] B. Figueiredo, A. Bifulco, C. Paiva, Ângela Maia, E. Fernandes, and R. Matos, *History of childhood abuse in portuguese parents,* Child Abuse & Neglect **28**, 669 (2004).

[212] P. Campos Pinto, *Out of the shadows: Violence against girls and women with disabilities in portugal,* Alter **10**, 137 (2016), interroger les sociétés contemporaines à la lumière du handicap. Quatrième conférence d'Alter 2015 / Questioning contemporary societies through the lens of disability. Fourth Alter conference 2015. Éditeurs / Guest editors : Noémie Rapegno, Isabelle Ville.

[213] S. Neves, M. Cameira, and S. Caridade, *Beliefs on marital violence among portuguese adolescents: Monitoring their evolution and relationship with dating violence,* Violence and Victims **36**, 363 (2021), https://connect.springerpub.com/content/sgrvv/36/3/363.full.pdf .

[214] S. L. Frazão, M. S. Silva, P. Norton, and T. Magalhães, *Domestic violence against elderly with disability,* Journal of Forensic and Legal Medicine **28**, 19 (2014).

[215] I. M. Caldas, A. C. Grams, A. Afonso, and T. Magalhães, *Oral injuries in victims involving intimate partner violence,* Forensic Science International **221**, 102 (2012).

[216] P. Vieira-Pinto, J. I. Muñoz-Barús, T. Taveira-Gomes, M. J. Vidal-Alves, and T. Magalhães, *Intimate partner violence against women. does violence decrease after the entry of the alleged offender into the criminal justice system?* Forensic sciences research , 1 (2021).

[217] A. G. Pinto, *A construção da identidade da mulher num acórdão sobre violência doméstica,* (2021).

[218] L. B. Cattaneo and L. A. Goodman, *Risk factors for reabuse in intimate partner violence: A cross-disciplinary critical review,* Trauma, Violence, & Abuse **6**, 141 (2005), pMID: 15753198, https://doi.org/10.1177/1524838005275088 .

[219] L. Barnham, G. C. Barnes, and L. W. Sherman, *Targeting escalation of intimate partner violence: Evidence from 52,000 offenders,* Cambridge Journal of Evidence-Based Policing **1**, 116 (2017).

[220] E. Sagi and M. Dehghani, *Measuring moral rhetoric in text,* Social science computer review **32**, 132 (2014).

[221] D. Oberski, *Mixture models: Latent profile and latent class analysis,* in *Modern statistical methods for HCI* (Springer, 2016) pp. 275–287.

[222] J. K. Vermunt and J. Magidson, *Latent class analysis,* The sage encyclopedia of social sciences research methods **2**, 549 (2004).

[223] E. Molin, P. Mokhtarian, and M. Kroesen, *Multimodal travel groups and attitudes: A latent class cluster analysis of dutch travelers,* Transportation Research Part A: Policy and Practice **83**, 14 (2016).

[224] K. L. Nylund, T. Asparouhov, and B. O. Muthén, *Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study,* Structural equation modeling: A multidisciplinary Journal **14**, 535 (2007).

[225] J. R. Silver and E. Silver, *Why are conservatives more punitive than liberals? a moral foundations approach.* Law and Human Behavior **41**, 258 (2017).

[226] M. Hengstler, E. Enkel, and S. Duelli, *Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices,* Technological Forecasting and Social Change **105**, 105 (2016).

[227] P. Lin, K. Abney, and R. Jenkins, *Robot ethics 2.0: from autonomous cars to artificial intelligence* (Oxford University Press, 2017).

[228] S. O.-R. A. V. S. Committee *et al.*, *Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems,* SAE Standard J **3016**, 1 (2014).

[229] R. M. Gandia, F. Antonialli, B. H. Cavazza, A. M. Neto, D. A. d. Lima, J. Y. Sugano, I. Nicolai, and A. L. Zambalde, *Autonomous vehicles: scientometric and bibliometric review,* Transport reviews **39**, 9 (2019).

[230] D. Danks and A. J. London, *Regulating autonomous systems: Beyond standards,* IEEE Intelligent Systems **32**, 88 (2017).

[231] J. M. Anderson, N. Kalra, K. D. Stanley, P. Sorensen, C. Samaras, and O. A. Oluwatola, *Autonomous vehicle technology: A guide for policymakers* (Rand Corporation, 2014).

[232] N. J. Goodall, *Can you program ethics into a self-driving car?* IEEE Spectrum **53**, 28 (2016).

[233] B. Lundgren, *Safety requirements vs. crashing ethically: what matters most for policies on autonomous vehicles,* AI & SOCIETY , 1 (2020).

[234] G. Keeling, *Why trolley problems matter for the ethics of automated vehicles,* Science and engineering ethics **26**, 293 (2020).

[235] K. Huang, J. D. Greene, and M. Bazerman, *Veil-of-ignorance reasoning favors the greater good,* Proceedings of the National Academy of Sciences **116**, 23989 (2019).

[236] G. Keeling, *Against leben's rawlsian collision algorithm for autonomous vehicles,* in *Philosophy and Theory of Artificial Intelligence 2017*, edited by V. C. Müller (Springer International Publishing, Cham, 2018) pp. 259–272.

[237] A. Hevelke and J. Nida-Rümelin, *Responsibility for crashes of autonomous vehicles: An ethical analysis,* Science and Engineering Ethics **21**, 619 (2015).

[238] L. Collingwood, *Privacy implications and liability issues of autonomous vehicles,* Information & Communications Technology Law **26**, 32 (2017), https://doi.org/10.1080/13600834.2017.1269871 .

[239] M. R. Endsley, *From here to autonomy: Lessons learned from human–automation research,* Human Factors **59**, 5 (2017), pMID: 28146676, https://doi.org/10.1177/0018720816681350 .

[240] M. Taiebat, A. L. Brown, H. R. Safford, S. Qu, and M. Xu, *A review on energy, environmental, and sustainability implications of connected and automated vehicles,* Environ. Sci. Technol. **52**, 11449 (2018).

[241] F. B. Aydemir and F. Dalpiaz, *A roadmap for ethics-aware software engineering,* in *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (2018) pp. 15–21.

[242] F. Santoni de Sio and J. van den Hoven, *Meaningful human control over autonomous systems: A philosophical account,* Frontiers in Robotics and AI **5**, 15 (2018).

[243] K. Kirkpatrick, *The moral challenges of driverless cars,* Communications of the ACM **58**, 19 (2015).

[244] D. J. Fagnant and K. Kockelman, *Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations,* Transportation Research Part A: Policy and Practice **77**, 167 (2015).

[245] J. Van den Hoven, P. Vermaas, and I. Van de Poel, *Handbook of ethics, values and technological design* (Springer, 2015).

[246] B. C. Soriano, S. L. Dougherty, B. G. Soublet, and K. J. Triepke, *Autonomous vehicles: A perspective from the california department of motor vehicles,* in *Road Vehicle Automation*, edited by G. Meyer and S. Beiker (Springer International Publishing, Cham, 2014) pp. 15–24.

[247] C. Menon and R. Alexander, *A safety-case approach to the ethics of autonomous vehicles,* Safety and Reliability **39**, 33 (2020), https://doi.org/10.1080/09617353.2019.1697918 .

[248] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher, *The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration,* Journal of clinical epidemiology **62**, e1 (2009).

[249] B.-A. Schuelke-Leech, S. R. Jordan, and B. Barry, *Regulating autonomy: An assessment of policy language for highly automated vehicles,* Review of Policy Research **36**, 547 (2019).

[250] B. Van Wee, *Transport and ethics: ethics and the evaluation of transport policies and projects* (Edward Elgar Publishing, 2011).

[251] J. J. Thomson, *The trolley problem,* Yale LJ **94**, 1395 (1984).

[252] P. Foot, *The problem of abortion and the doctrine of the double effect,* Oxford Review **5**, 5 (1967).

[253] N. Kallioinen, M. Pershina, J. Zeiser, F. Nosrat Nezami, A. Stephan, G. Pipa, and P. König, *Moral judgements on the actions of self-driving cars and human drivers in dilemma situations from different perspectives,* Frontiers in psychology **10**, 2415 (2019).

[254] T. Morita and S. Managi, *Autonomous vehicles: Willingness to pay and the social dilemma,* Transportation Research Part C: Emerging Technologies **119**, 102748 (2020).

[255] J. Harris, *The immoral machine,* CAMBRIDGE QUARTERLY OF HEALTHCARE ETHICS **29**, 71 (2020).

[256] C. Luetge, *The german ethics code for automated and connected driving,* Philosophy & Technology **30**, 547 (2017).

[257] B. Huang, S. van Cranenburgh, and C. G. Chorus, *Death by automation,* European Journal of Transport and Infrastructure Research **20**, 71 (2020).

[258] N. Goodall, *More than trolley,* Transfers **9**, 45 (2019).

[259] G. Keeling, K. Evans, S. M. Thornton, G. Mecacci, and F. S. de Sio, *Four perspectives on what matters for the ethics of automated vehicles,* in *Automated Vehicles Symposium* (Springer, 2019) pp. 49–60.

[260] A. Brown, G. Rodriguez, B. Best, K. T. Hoang, H. Safford, G. Anderson, and M. C. D'Agostino, *Federal, state, and local governance of automated vehicles,* Institute of Transportation Studies & Policy Institute for Energy, Environment and the Economy, University of California Davis (2018).

[261] N. Kalra and S. M. Paddock, *Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?* Transportation Research Part A: Policy and Practice **94**, 182 (2016).

[262] S. Parkinson, P. Ward, K. Wilson, and J. Miller, *Cyber threats facing autonomous and connected vehicles: Future challenges,* IEEE transactions on intelligent transportation systems **18**, 2898 (2017).

[263] R. Sparrow and M. Howard, *When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport,* Transportation Research Part C: Emerging Technologies **80**, 206 (2017).

[264] J. Borenstein, J. R. Herkert, and K. W. Miller, *Self-driving cars and engineering ethics: the need for a system level analysis,* Science and engineering ethics **25**, 383 (2019).

[265] S. Nyholm and J. Smids, *The ethics of accident-algorithms for self-driving cars: An applied trolley problem?* Ethical theory and moral practice **19**, 1275 (2016).

[266] B. H. Fried, *What does matter? the case for killing the trolley problem (or letting it die),* The Philosophical Quarterly **62**, 505 (2012).

[267] S. A. Bagloee, M. Tavana, M. Asadi, and T. Oliver, *Autonomous vehicles: challenges, opportunities, and future implications for transportation policies,* Journal of modern transportation **24**, 284 (2016).

[268] N. J. Goodall, *Ethical decision making during automated vehicle crashes,* Transportation Research Record **2424**, 58 (2014).

[269] S. S. Wu, *Autonomous vehicles, trolley problems, and the law,* Ethics and Information Technology , 1 (2019).

[270] G. Keeling, *Against leben's rawlsian collision algorithm for autonomous vehicles,* in *3rd Conference on" Philosophy and Theory of Artificial Intelligence* (Springer, 2017) pp. 259–272.

[271] R. Noothigattu, S. S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. D. Procaccia, *A voting-based system for ethical decision making,* in *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[272] H. Etienne, *When ai ethics goes astray: A case study of autonomous vehicles,* Social Science Computer Review , 0894439320906508 (2020).

[273] S. Marakby, *A Matter of Trust: Ford's Approach to Developing Self-Driving Vehicles*, Tech. Rep. (Ford Autonomous Vehicles LLC, 2018).

[274] J. Weast, M. Yurdana, and A. Jordan, *A Matter of Trust: How Smart Design Can Accelerate Automated Vehicle Adoption*, Tech. Rep. (Intel, 2016).

[275] Daimler, *Reinventing Safety: A Joint Approach to Automated Driving Systems*, Tech. Rep. (Mercedes-Benz and Bosch, 2018).

[276] Valeo, *Meet the Future 2016 Activity and Sustainable Development Report*, Tech. Rep. (Valeo, 2016).

[277] Mando, *Mando Sustainability Report 2018*, Tech. Rep. (Mando Corporation, 2018).

[278] Toyota, *Automated Driving at Toyota: Vision, Strategy and Development*, Tech. Rep. (Toyota Motor Corporation, n.d.).

[279] Nuro, *Delivering Safety: Nuro's Approach*, Tech. Rep. (Nuro, 2018).

[280] N. J. Goodall, *Machine ethics and automated vehicles,* in *Road vehicle automation* (Springer, 2014) pp. 93–102.

[281] GM, *Self-Driving Safety Report*, Tech. Rep. (General Motors, 2018).

[282] Nvidia, *Nvidia Self-Driving Safety Report*, Tech. Rep. (Nvidia, 2018).

[283] AutoX, *The Autox Safety Factor*, Tech. Rep. (Auto X, 2018).

[284] Mobileye, *Implementing the RSS Model on NHTSA Pre-Crash Scenarios*, Tech. Rep. (Intel Corp, n.d.).

[285] D. D. Heikoop, M. Hagenzieker, G. Mecacci, S. Calvert, F. Santoni De Sio, and B. van Arem, *Human behaviour with automated driving systems: a quantitative framework for meaningful human control,* Theoretical issues in ergonomics science **20**, 711 (2019).

[286] J. M. Fischer and M. Ravizza, *Responsibility and control: A theory of moral responsibility* (Cambridge university press, 2000).

[287] A. Matthias, *The responsibility gap: Ascribing responsibility for the actions of learning automata,* Ethics and information technology **6**, 175 (2004).

[288] S. Nyholm, *The ethics of crashes with self-driving cars: A roadmap, i,* Philosophy Compass **13**, e12507 (2018).

[289] I. Van de Poel, J. N. Fahlquist, N. Doorn, S. Zwart, and L. Royakkers, *The problem of many hands: Climate change as an example,* Science and engineering ethics **18**, 49 (2012).

[290] Uber, *A Principled Approach To Safety*, Tech. Rep. (Uber, 2018).

[291] Zoox, *Safety innovation at Zoox: Setting the bar for safety in autonomous mobility*, Tech. Rep. (Zoox, 2018).

[292] M. Wood, P. Robbel, M. Maass, R. D. Tebbens, M. Meijs, M. Harb, J. Reach, and K. Robinson, *Safety First for Automated Driving*, Tech. Rep. (Mercedes-Benz, Daimler, Intel Corp, 2019).

[293] T. Csizmadia, *Ensuring Safe Self-Driving AImotive's Development Puts Safety First*, Tech. Rep. (AImotive, 2018).

[294] R. Mulgan, *'accountability': An ever-expanding concept?* Public administration **78**, 555 (2000).

[295] M. Coeckelbergh, *Artificial intelligence, responsibility attribution, and a relational justification of explainability,* Science and Engineering Ethics **26**, 2051 (2020).

[296] M. Coeckelbergh, *Responsibility and the moral phenomenology of using self-driving cars,* Applied Artificial Intelligence **30**, 748 (2016).

[297] J. Borenstein, J. Herkert, and K. Miller, *Self-driving cars: Ethical responsibilities of design engineers,* IEEE Technology and Society Magazine **36**, 67 (2017).

[298] C. Misselhorn, *Collective agency and cooperation in natural and artificial systems,* in *Collective Agency and Cooperation in Natural and Artificial Systems* (Springer, 2015) pp. 3–24.

[299] G. E. Marchant and R. A. Lindor, *The coming collision between autonomous vehicles and the liability system,* Santa Clara L. Rev. **52**, 1321 (2012).

[300] J. K. Gurney, *Sue my car not me: Products liability and accidents involving autonomous vehicles,* U. Ill. JL Tech. & Pol'y , 247 (2013).

[301] J. K. Gurney, *Crashing into the unknown: An examination of crash-optimization algorithms through the two lanes of ethics and law,* Alb. L. Rev. **79**, 183 (2015).

[302] K. Funkhouser, *Paving the road ahead: autonomous vehicles, products liability, and the need for a new approach,* Utah L. Rev. , 437 (2013).

[303] A. P. Garza, *Look ma, no hands: wrinkles and wrecks in the age of autonomous vehicles,* New Eng. L. Rev. **46**, 581 (2011).

[304] A. Taeihagh and H. S. M. Lim, *Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks,* Transport reviews **39**, 103 (2019).

[305] Honda, *Honda Sustainability Report*, Tech. Rep. (Honda, 2015).

[306] BMW, *Sustainable Value Report*, Tech. Rep. (BMW, 2016).

[307] A. Shashua and S. Shalev-Shwartz, *A Plan to Develop Autonomous Vehicles. And Prove it.*, Tech. Rep. (Intel, 2017).

[308] Nissan, *Financial Information as of March 31, 2017*, Tech. Rep. (Nissan, 2017).

[309] K.-H. Yu, A. L. Beam, and I. S. Kohane, *Artificial intelligence in healthcare,* Nature biomedical engineering **2**, 719 (2018).

[310] V. Kaul, S. Enslin, and S. A. Gross, *History of artificial intelligence in medicine,* Gastrointestinal Endoscopy **92**, 807 (2020).

[311] C. Wallis, *How artificial intelligence will change medicine.* Nature **576**, S48 (2019).

[312] M. van Hartskamp, S. Consoli, W. Verhaegh, M. Petkovic, and A. van de Stolpe, *Artificial intelligence in clinical health care applications,* Interactive Journal of Medical Research **8**, e12100 (2019).

[313] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, *Artificial intelligence in healthcare: past, present and future,* Stroke and vascular neurology **2**, 230 (2017).

[314] J. Wolff, J. Pauling, A. Keck, and J. Baumbach, *The economic impact of artificial intelligence in health care: Systematic review,* J Med Internet Res **22**, e16866 (2020).

[315] W. Raghupathi and V. Raghupathi, *Big data analytics in healthcare: promise and potential,* Health Information Science and Systems **2**, 3 (2014).

[316] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, *The practical implementation of artificial intelligence technologies in medicine,* Nature Medicine **25**, 30 (2019).

[317] J. Morley, C. C. Machado, C. Burr, J. Cowls, I. Joshi, M. Taddeo, and L. Floridi, *The ethics of ai in health care: A mapping review,* Social Science & Medicine **260**, 113172 (2020).

[318] M. J. Rigby, *Ethical dimensions of using artificial intelligence in health care,* AMA Journal of Ethics **21**, 121 (2019).

[319] W. N. Price and I. G. Cohen, *Privacy in the age of medical big data,* Nature medicine **25**, 37 (2019).

[320] E. Crigger and C. Khoury, *Making policy on augmented intelligence in health care,* AMA journal of ethics **21**, 188 (2019).

[321] D. P. Dos Santos, D. Giese, S. Brodehl, S. Chon, W. Staab, R. Kleinert, D. Maintz, and B. Baeßler, *Medical students' attitude towards artificial intelligence: a multicentre survey,* European radiology **29**, 1640 (2019).

[322] J. van Hoek, A. Huber, A. Leichtle, K. Härmä, D. Hilt, H. von Tengg-Kobligk, J. Heverhagen, and A. Poellinger, *A survey on the future of radiology among radiologists, medical students and surgeons: Students and surgeons tend to be more skeptical about artificial intelligence and radiologists may fear that other disciplines take over,* European Journal of Radiology **121**, 108742 (2019).

[323] C. Sit, R. Srinivasan, A. Amlani, K. Muthuswamy, A. Azam, L. Monzon, and D. S. Poon, *Attitudes and perceptions of uk medical students towards artificial intelligence and radiology: a multicentre survey,* Insights into Imaging **11**, 14 (2020).

[324] S. K. G. Ooi, A. Makmur, A. Y. Q. Soon, S. Fook-Chong, C. Liew, S. Y. Sia, Y. H. Ting, and C. Y. Lim, *Attitudes toward artificial intelligence in radiology with learner needs assessment within radiology residency programmes: a national multi-programme survey,* Singapore Med J **1**, 22 (2019).

[325] C. J. Park, P. H. Yi, and E. L. Siegel, *Medical student perspectives on the impact of artificial intelligence on the practice of medicine,* Current Problems in Diagnostic Radiology (2020), https://doi.org/10.1067/j.cpradiol.2020.06.011.

[326] F. Jungmann, T. Jorg, F. Hahn, D. Pinto dos Santos, S. M. Jungmann, C. Düber, P. Mildenberger, and R. Kloeckner, *Attitudes toward artificial intelligence among radiologists, it specialists, and industry,* Academic Radiology (2020), https://doi.org/10.1016/j.acra.2020.04.011.

[327] F. Coppola, L. Faggioni, D. Regge, A. Giovagnoni, R. Golfieri, C. Bibbolino, V. Miele, E. Neri, and R. Grassi, *Artificial intelligence: radiologists' expectations and opinions gleaned from a nationwide online survey,* La radiologia medica (2020).

[328] S. Sarwar, A. Dent, K. Faust, M. Richer, U. Djuric, R. Van Ommeren, and P. Diamandis, *Physician perspectives on integration of artificial intelligence into diagnostic pathology,* npj Digital Medicine **2**, 28 (2019).

[329] S. Polesie, M. Gillstedt, H. Kittler, A. Lallas, P. Tschandl, I. Zalaudek, and J. Paoli, *Attitudes towards artificial intelligence within dermatology: an international online survey,* British Journal of Dermatology , 159 (2020).

[330] C. Shen, C. Li, F. Xu, Z. Wang, X. Shen, J. Gao, R. Ko, Y. Jing, X. Tang, R. Yu, *et al.*, *Web-based study on chinese dermatologists' attitudes towards artificial intelligence,* Annals of Translational Medicine **8** (2020).

[331] K. S. Chan and N. Zary, *Applications and challenges of implementing artificial intelligence in medical education: Integrative review,* JMIR Med Educ **5**, e13930 (2019).

[332] A. G. van der Niet and A. Bleakley, *Where medical education meets artificial intelligence: 'does technology care?',* Medical Education **55**, 30 (2021), https://onlinelibrary.wiley.com/doi/pdf/10.1111/medu.14131 .

[333] V. B. Kolachalama and P. S. Garg, *Machine learning and medical education,* npj Digital Medicine **1**, 54 (2018).

[334] S. Gao, L. He, Y. Chen, D. Li, and K. Lai, *Public perception of artificial intelligence in medical care: Content analysis of social media,* Journal of Medical Internet Research **22**, e16649 (2020).

[335] P. M. Doraiswamy, C. Blease, and K. Bodner, *Artificial intelligence and the future of psychiatry: Insights from a global physician survey,* Artificial Intelligence in Medicine **102**, 101753 (2020).

[336] N. Brault and M. Saxena, *For a critical appraisal of artificial intelligence in healthcare: The problem of bias in mhealth,* Journal of Evaluation in Clinical Practice **n/a** (2020), https://doi.org/10.1111/jep.13528, https://onlinelibrary.wiley.com/doi/pdf/10.1111/jep.13528 .

[337] E. Tat, D. L. Bhatt, and M. G. Rabbat, *Addressing bias: artificial intelligence in cardiovascular medicine,* The Lancet Digital Health **2**, e635 (2020).

[338] D. Cirillo, S. Catuara-Solarz, C. Morey, E. Guney, L. Subirats, S. Mellino, A. Gigante, A. Valencia, M. J. Rementeria, A. S. Chadha, and N. Mavridis, *Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare,* npj Digital Medicine **3**, 81 (2020).

[339] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, *Key challenges for delivering clinical impact with artificial intelligence,* BMC medicine **17**, 195 (2019).

[340] A. Vellido, *Societal issues concerning the application of artificial intelligence in medicine,* (2019).

[341] B. McKeown and D. B. Thomas, *Q methodology*, Vol. 66 (Sage publications, 2013).

[342] S. Watts and P. Stenner, *Doing q methodology: theory, method and interpretation,* Qualitative research in psychology **2**, 67 (2005).

[343] S. Watts and P. Stenner, *Doing Q methodological research: Theory, method & interpretation* (Sage, 2012).

[344] W. Stephenson, *The study of behavior; Q-technique and its methodology.* (University of Chicago Press, Chicago, IL, US, 1953) pp. ix, 376–ix, 376.

[345] W. Stephenson, *Technique of factor analysis,* Nature **136**, 297 (1935).

[346] A. Zabala, C. Sandbrook, and N. Mukherjee, *When and how to use q methodology to understand perspectives in conservation research,* Conservation Biology **32**, 1185 (2018).

[347] S. Kamal, M. Kocór, and M. Grodzińska-Jurczak, *Quantifying human subjectivity using q method: When quality meets quantity.* Qualitative Sociology Review **10** (2014).

[348] H. A. Linstone, M. Turoff, *et al.*, *The delphi method* (Addison-Wesley Reading, MA, 1975).

[349] W. Stephenson, *Concourse theory of communication,* Communication **3**, 21 (1978).

[350] S. Banasick, *Ken-q analysis (version 1.0.6),* (2019).

[351] P. Schmolck, *Pq-method, version 2.11 manual,* Neibiderg, Germany: University (2002).

[352] M. Ringnér, *What is principal component analysis?* Nature Biotechnology **26**, 303 (2008).

[353] S. Wold, K. Esbensen, and P. Geladi, *Principal component analysis,* Chemometrics and Intelligent Laboratory Systems **2**, 37 (1987), proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[354] N. Akhtar-Danesh *et al.*, *A comparison between major factor extraction and factor rotation techniques in q-methodology,* Open Journal of Applied Sciences **7**, 147 (2017).

[355] A. Zabala and U. Pascual, *Bootstrapping q methodology to improve the understanding of human perspectives,* PloS one **11**, e0148087 (2016).

[356] R. I. Field, *Why is health care regulation so complex?* P & T : a peer-reviewed journal for formulary management **33**, 607 (2008).

[357] T. Minssen, S. Gerke, M. Aboy, N. Price, and G. Cohen, *Regulatory responses to medical machine learning,* Journal of Law and the Biosciences (2020).

[358] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, *Man is to computer programmer as woman is to homemaker? debiasing word embeddings,* in *Advances in neural information processing systems* (2016) pp. 4349–4357.

[359] A. Kadambi, *Achieving fairness in medical devices,* Science **372**, 30 (2021).

[360] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, *Explaining explanations: An overview of interpretability of machine learning,* in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (IEEE, 2018) pp. 80–89.

[361] S. Gerke, B. Babic, T. Evgeniou, and I. G. Cohen, *The need for a system view to regulate artificial intelligence/machine learning-based software as medical device,* npj Digital Medicine **3**, 53 (2020).

[362] J.-A. Cervantes, S. López, L.-F. Rodríguez, S. Cervantes, F. Cervantes, and F. Ramos, *Artificial moral agents: A survey of the current status,* Science and engineering ethics , 1 (2019).

[363] W. A. Bauer, *Virtuous vs. utilitarian artificial moral agents,* AI & SOCIETY **35**, 263 (2020).

[364] P. Formosa and M. Ryan, *Making moral machines: why we need artificial moral agents,* AI & SOCIETY (2020).

[365] M. Hunyadi, *Artificial moral agents. really?* in *Wording Robotics: Discourses and Representations on Robotics*, edited by J.-P. Laumond, E. Danblon, and C. Pieters (Springer International Publishing, Cham, 2019) pp. 59–69.

[366] A. Poulsen, M. Anderson, S. L. Anderson, B. Byford, F. Fossa, E. L. Neely, A. Rosas, and A. Winfield, *Responses to a critique of artificial moral agents,* arXiv preprint arXiv:1903.07021 (2019).

[367] M. Anderson and S. L. Anderson, *Geneth: A general ethical dilemma analyzer,* Paladyn, Journal of Behavioral Robotics **9**, 337 (2018).

[368] D. Vanderelst and A. Winfield, *An architecture for ethical robots inspired by the simulation theory of cognition,* Cognitive Systems Research **48**, 56 (2018).

[369] D. Howard and I. Muntean, *Artificial moral cognition: moral functionalism and autonomous moral agency,* in *Philosophy and computing* (Springer, 2017) pp. 121–159.

[370] M. Dehghani, E. Tomai, K. Forbus, R. Iliev, and M. Klenk, *Moraldm: A computational modal of moral decision-making,* in *Proceedings of the 30th Annual Conference of the Cognitive Science Society (CogSci)* (Citeseer, 2008).

[371] D. Behdadi and C. Munthe, *A normative approach to artificial moral agency,* MINDS AND MACHINES (2020).

[372] R. V. Yampolskiy, *Artificial intelligence safety engineering: Why machine ethics is a wrong approach,* in *Philosophy and theory of artificial intelligence* (Springer, 2013) pp. 389–396.

[373] W. Wallach, C. Allen, and I. Smit, *Machine morality: bottom-up and top-down approaches for modelling human moral faculties,* Ai & Society **22**, 565 (2008).

[374] J. P. Sullins, *Artificial moral agency in technoethics,* in *Handbook of research on technoethics* (IGI Global, 2009) pp. 205–221.

[375] F. Fossa, *Artificial moral agents: moral mentors or sensible tools?* Ethics and Information Technology **20**, 115 (2018).

[376] V. C. Müller and N. Bostrom, *Future progress in artificial intelligence: A survey of expert opinion,* in *Fundamental issues of artificial intelligence* (Springer, 2016) pp. 555–572.

[377] K. Sotala and R. V. Yampolskiy, *Responses to catastrophic agi risk: a survey,* Physica Scripta **90**, 018001 (2014).

[378] D. Behdadi and C. Munthe, *A normative approach to artificial moral agency,* Minds and Machines **30**, 195 (2020).

[379] M. Anderson and S. L. Anderson, *Robot be good,* Scientific American **303**, 72 (2010).

[380] M. Scheutz, *The need for moral competency in autonomous agent architectures,* in *Fundamental issues of artificial intelligence* (Springer, 2016) pp. 517–527.

[381] M. Anderson and S. L. Anderson, *Machine ethics: Creating an ethical intelligent agent,* AI magazine **28**, 15 (2007).

[382] S. L. Anderson, *Machine metaethics,* Machine ethics. Cambridge University Press, Cambridge , 21 (2011).

[383] S. L. Anderson and M. Anderson, *Ai and ethics,* AI and Ethics **1**, 27 (2021).

[384] R. Tonkens, *Out of character: on the creation of virtuous machines,* Ethics and Information Technology **14**, 137 (2012).

[385] R. Tonkens, *A challenge for machine ethics,* Minds and Machines **19**, 421 (2009).

[386] D. Vanderelst and A. Winfield, *The dark side of ethical robots,* in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018) pp. 317–322.

[387] U. Köse, *Are we safe enough in the future of artificial intelligence? a discussion on machine ethics and artificial intelligence safety,* BRAIN. Broad Research in Artificial Intelligence and Neuroscience **9**, 184 (2018).

[388] M. R. Waser, *Designing, implementing and enforcing a coherent system of laws, ethics and morals for intelligent machines (including humans),* Procedia Computer Science **71**, 106 (2015).

[389] B. A. Mabaso, *Computationally rational agents can be moral agents,* Ethics and Information Technology (2020), 10.1007/s10676-020-09527-1.

[390] T. M. Powers, *Prospects for a kantian machine,* IEEE Intelligent Systems **21**, 46 (2006).

[391] T. M. Powers, *Machines and moral reasoning,* Philosophy Now **72**, 15 (2009).

[392] J. N. Hooker and T. W. N. Kim, *Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic,* in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018) pp. 130–136.

[393] S. L. Anderson and M. Anderson, *A prima facie duty approach to machine ethics: Machine learning of features of ethical dilemmas, prima facie duties, and decision principles through a dialogue with ethicists,* Machine ethics , 476 (2011).

[394] C. Shulman, H. Jonsson, and N. Tarleton, *Which consequentialism? machine ethics and moral divergence,* in *Asia-Pacific Conference on Computing and Philosophy (APCAP 2009), Tokyo, Japan* (Citeseer, 2009).

[395] N.-M. Aliman and L. Kester, *Augmented utilitarianism for agi safety,* in *International Conference on Artificial General Intelligence* (Springer, 2019) pp. 11–21.

[396] P. Gamez, D. B. Shank, C. Arnold, and M. North, *Artificial virtue: the machine question and perceptions of moral character in artificial moral agents,* AI & Society (2020).

[397] N. S. Govindarajulu, S. Bringsjord, R. Ghosh, and V. Sarathy, *Toward the engineering of virtuous machines,* in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019) pp. 29–35.

[398] K. Bogosian, *Implementation of moral uncertainty in intelligent machines,* Minds and Machines **27**, 591 (2017).

[399] E. Awad, M. Anderson, S. L. Anderson, and B. Liao, *An approach for combining ethical principles with public opinion to guide public policy,* Artificial Intelligence **287**, 103349 (2020).

[400] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, and A. Bernstein, *Implementations in machine ethics: A survey,* ACM Comput. Surv. **53** (2021), 10.1145/3419633.

[401] M. Brundage, *Limitations and risks of machine ethics,* Journal of Experimental & Theoretical Artificial Intelligence **26**, 355 (2014).

[402] B. F. Malle, *Integrating robot ethics and machine morality: the study and design of moral competence in robots,* Ethics and Information Technology **18**, 243 (2016).

[403] V. Nallur, *Landscape of machine implemented ethics,* Science and Engineering Ethics **26**, 2381 (2020).

[404] C. Misselhorn, *Artificial morality. concepts, issues and challenges,* Society **55**, 161 (2018).

[405] E. Dogan, R. Chatila, S. Chauvier, K. Evans, P. Hadjixenophontos, and J. Perrin, *Ethics in the design of automated vehicles: The avethics project.* in *EDIA@ ECAI* (2016) pp. 10–13.

[406] B. Liao, M. Anderson, and S. L. Anderson, *Representation, justification, and explanation in a value-driven agent: an argumentation-based approach,* AI and Ethics , 1 (2020).

[407] P. Bremner, L. A. Dennis, M. Fisher, and A. F. Winfield, *On proactive, transparent, and verifiable ethical reasoning for robots,* Proceedings of the IEEE **107**, 541 (2019).

[408] N. P. Shaw, A. Stöckel, R. W. Orr, T. F. Lidbetter, and R. Cohen, *Towards provably moral ai agents in bottom-up learning frameworks,* in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018) pp. 271–277.

[409] K. E. Himma, *Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?* Ethics and Information Technology **11**, 19 (2009).

[410] A. Fritz, W. Brandt, H. Gimpel, and S. Bayer, *Moral agency without responsibility? analysis of three ethical models of human-computer interaction in times of artificial intelligence (ai),* De Ethica **6**, 3 (2020).

[411] F. S. Grodzinsky, K. W. Miller, and M. J. Wolf, *The ethics of designing artificial agents,* Ethics and Information Technology **10**, 115 (2008).

[412] C. Misselhorn, *Artificial systems with moral capacities? a research design and its implementation in a geriatric care system,* Artificial Intelligence **278**, 103179 (2020).

[413] J. Hallamaa and T. Kalliokoski, *How ai systems challenge the conditions of moral agency?* in *International Conference on Human-Computer Interaction* (Springer, 2020) pp. 54–64.

[414] L. Floridi and J. W. Sanders, *On the morality of artificial agents,* Minds and Machines **14**, 349 (2004).

[415] M. Talbert, *Moral responsibility: an introduction* (John Wiley & Sons, 2016).

[416] J. Danaher, *The rise of the robots and the crisis of moral patiency,* AI & SOCIETY **34**, 129 (2019).

[417] C. J. Headleand, W. J. Teahan, and L. ap Cenydd, *Sexbots: a case for artificial ethical agents,* Connection Science , 1 (2019).

[418] A. Poulsen and O. K. Burmeister, *Overcoming carer shortages with care robots: Dynamic value trade-offs in run-time,* Australasian Journal of Information Systems **23** (2019).

[419] I. Verdiesen, F. Santoni de Sio, and V. Dignum, *Accountability and control over autonomous weapon systems: A framework for comprehensive human oversight,* Minds and Machines **31**, 137 (2021).

[420] N. Bostrom, *How long before superintelligence?* International Journal of Futures Studies **2** (1998).

[421] N. Bostrom and E. Yudkowsky, *The ethics of artificial intelligence,* The Cambridge handbook of artificial intelligence **1**, 316 (2014).

[422] S. Russell, *It's not too soon to be wary of ai: We need to act now to protect humanity from future superintelligent machines,* IEEE Spectrum **56**, 46 (2019).

[423] S. Russell and J. Bohannon, *Artificial intelligence. fears of an ai pioneer,* Science (New York, NY) **349**, 252 (2015).

[424] V. Dignum, *Responsible artificial intelligence: Designing ai for human values,* Discoveries **1**, 1 (2017).

[425] A. Feller, E. Pierson, S. Corbett-Davies, and S. Goel, *A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear,* The Washington Post (2016).

[426] R. Wexler, *When a computer program keeps you in jail: How computers are harming criminal justice,* New York Times **13** (2017).

[427] C. Allen, G. Varner, and J. Zinser, *Prolegomena to any future artificial moral agent,* Journal of Experimental & Theoretical Artificial Intelligence **12**, 251 (2000).

[428] R. Dobbe, T. K. Gilbert, and Y. Mintz, *Hard choices in artificial intelligence: Addressing normative uncertainty through sociotechnical commitments,* arXiv preprint arXiv:1911.09005 (2019).

[429] W. MacAskill, *Normative uncertainty*, Ph.D. thesis, University of Oxford (2014).

[430] W. MacAskill, *Normative uncertainty as a voting problem,* Mind **125**, 967 (2016).

[431] W. MacAskill, K. Bykvist, and T. Ord, *Moral Uncertainty* (Oxford University Press, 2020).

[432] I. Nissan-Rozen, *Against moral hedging,* Economics & Philosophy **31**, 349 (2015).

[433] T. Żuradzki, *Meta-reasoning in making moral decisions under normative uncertainty,* in *2016). Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, Vol. 2 (2015) pp. 1093–1104.

[434] H. Zhao, K. Dimovitz, B. Staveland, and L. Medsker, *Responding to challenges in the design of moral autonomous vehicles,* in *The 2016 AAAI Fall Symposium Series: Cognitive Assistance in Government and Public Sector Applications, Technical Report FS-16-02* (2016) pp. 169–173.

[435] A. K. Faulhaber, A. Dittmer, F. Blind, M. A. Wächter, S. Timm, L. R. Sütfeld, A. Stephan, G. Pipa, and P. König, *Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles,* Science and engineering ethics , 1 (2018).

[436] J. Harris, *The immoral machine,* CAMBRIDGE QUARTERLY OF HEALTHCARE ETHICS **29**, 71 (2020).

[437] L. Floridi and J. W. Sanders, *On the morality of artificial agents,* Minds and machines **14**, 349 (2004).

[438] T. Lockhart, *Moral uncertainty and its consequences* (Oxford University Press, 2000).

[439] I. van de Poel, *Embedding values in artificial intelligence (ai) systems,* Minds and Machines , 1 (2020).

[440] M. Klenk, *How do technological artefacts embody moral values?* Philosophy & Technology **34**, 525 (2021).

[441] A. Backlund, *The definition of system,* Kybernetes **29**, 444 (2000).

[442] P. A. Samuelson, *A note on the pure theory of consumer's behaviour,* Economica **5**, 61 (1938).

[443] P. A. Samuelson, *Consumption theory in terms of revealed preference,* Economica **15**, 243 (1948).

[444] D. McFadden *et al.*, *Conditional logit analysis of qualitative choice behavior,* (1973).

[445] K. E. Train, *Discrete choice methods with simulation* (Cambridge university press, 2009).

[446] M. E. Ben-Akiva, S. R. Lerman, and S. R. Lerman, *Discrete choice analysis: theory and application to travel demand*, Vol. 9 (MIT press, 1985).

[447] C. F. Manski, *The structure of random utility models,* Theory and decision **8**, 229 (1977).

[448] J. Walker and M. Ben-Akiva, *Generalized random utility model,* Mathematical social sciences **43**, 303 (2002).

[449] H. Azari, D. Parks, and L. Xia, *Random utility theory for social choice,* in *Advances in Neural Information Processing Systems* (2012) pp. 126–134.

[450] W. H. Greene and D. A. Hensher, *A latent class model for discrete choice analysis: contrasts with mixed logit,* Transportation Research Part B: Methodological **37**, 681 (2003).

[451] J. Magidson, T. Eagle, and J. K. Vermunt, *New developments in latent class choice models,* in *Sawtooth Software Conference Proceedings* (2003) pp. 89–112.

[452] J. Magidson and J. K. Vermunt, *Latent class models,* The Sage handbook of quantitative methodology for the social sciences , 175 (2004).

[453] M. Kroesen, *Modeling the behavioral determinants of travel behavior: An application of latent transition analysis,* Transportation Research Part A: Policy and Practice **65**, 56 (2014).

[454] Y. Araghi, M. Kroesen, E. Molin, and B. Van Wee, *Revealing heterogeneity in air travelers' responses to passenger-oriented environmental policies: A discrete-choice latent class model,* International Journal of Sustainable Transportation **10**, 765 (2016).

[455] K. J. Lancaster, *A new approach to consumer theory,* Journal of political economy **74**, 132 (1966).

[456] C. Chorus, N. Mouter, and B. Pudane, *A taboo trade off model for discrete choice analysis,* in *International Choice Modelling Conference 2017* (2017).

[457] C. G. Chorus, *A new model of random regret minimization,* European Journal of Transport and Infrastructure Research **10** (2010).

[458] J. J. Louviere, D. A. Hensher, and J. D. Swait, *Stated choice methods: analysis and applications* (Cambridge university press, 2000).

[459] R. Shafer-Landau, *Ethical theory: an anthology*, Vol. 13 (John Wiley & Sons, 2012).

[460] A. A. Neath and J. E. Cavanaugh, *The bayesian information criterion: background, derivation, and applications,* Wiley Interdisciplinary Reviews: Computational Statistics **4**, 199 (2012).

[461] E. Commission, *White paper on artificial intelligence: A european approach to excellence and trust,* Com (2020) 65 Final (2020).

[462] H. H. Harman, *Modern factor analysis* (University of Chicago press, 1976).

[463] L. L. Thurstone, *Multiple factor analysis.* Psychological review **38**, 406 (1931).

A

Supplementary Information of Chapter 4

This Supplementary Information features the comments associated with the defining sorts of each factor in the q-methodological study - these anonymous comments concern the statements that participants ranked highest and lowest in the sorting step of the study.

A.1. Comments of Participants that loaded significantly on factors

A.1.1. Factor 1

sort 08140708

comment33: (s33) In a helthcare system all the decision should be advise by a medical doctor who understands the clinical issues and patients/services necessities rather than only worries about the financial aspects of healthcare management.comment26: (s26) A patient is much more than his physical problems. His symptoms are a result of his biopsychosocial components and his social environment is crucial in the management of his health problems. comment14: (s14) AI will help doctors making clinical choices, but patient-physician relationships are based also in trust, human expressions and empathy. comment18: (s18) In my opinion, AI will only help the physicians resolving their clinical doubts, but the last decision should never be given by AI.

sort 08150213

comment36: (s36) That is the main aim! To let physicians do what they were trained for - medicine - and alleviate many of the potentially automatic and time-consuming processes they have to daily face. The same rational is applied here for other healthcare professionals. comment27: (s27) AI and conventional medical practice should be synergistic. In the future, one should not replace the other. Instead, AI should aid physicians' decisions in daily practice. It is therefore unethical if a doctor is depriving his patients from having the best care possible by not complementing his practice with AI. comment16: (s16) In my

perspective, the role of a skilled physician is to take into consideration what a machine / AI tells him and make the correct connection with clinical reality. It is essentially to see if a certain AI result is clinically applicable and makes sense for each individual case. The MRIs are already reported by doctors (radiologists). In some centers MRIs can have an AI reading, but it is always a doctor who has to validate their results. Therefore, if a physician other than radiologist relies on MRI results, they are trusting in what another colleague has written and not what a machine has said. comment38: (s38) Many variables have to taken into consideration in medicine, that in particular situations I reckon can be very difficult to automate (social-demographic factors, all the background history, physical examination, etc).

sort 09091040
comment36: (s36) Less time needed for boring work means more time for challenging work. comment33: (s33) They define the problem and set the gold standard and can explain how AI tools should be designed to hep them best. comment18: (s18) No way there will always be human supervision. comment25: (s25) that is BS.

sort 09100943
comment19: (s19) AI can improve decision making. comment33: (s33) Professionals in the lead. comment22: (s22) Some kinds of routine work is already automated in hospital. comment11: (s11) The process of decision making by AI is in essence incomprehensible.

sort 09101106
comment19: (s19) I think AI will give MDs additional tools, like we've been using for ages. Someone to combine modalities and to put the results into the right (patient) context will always be needed. comment33: (s33) AI designers know the technology, but need MDs to design relevant products. comment18: (s18) As I mentioned above, I see AI as an additional tool, not as something that will replace MDs or decrease autonomy and authority. A MD will always have the final verdict. comment38: (s38) From my experience many clinical tasks are rather complex and thereby difficult to operationalize for a machine.

sort 09110126
comment37: (s37) AI already has such a broad range of tools and I have no doubt that most areas can benefit from AI if the possibilities are explained to the employees and they are able to give their input to the developers. comment33: (s33) The goal of a certain AI tool has to be defined together with the medical doctors to ensure clinical relevance. Besides that, the most important outcome of the tool (e.g. sensitivity or negative predictive value) has to be defined with them to ensure proper finetuning of parameters. Medical doctors are also able to give feedback on usability and possible errors. It is important that the medical doctors know how to use an AI tool for which particular goal to avoid usage of the model in an unsuitable situation. In my opinion, interpretation of the tool by medical doctors has ALWAYS to be investigated and aligned with the goal of the tool. comment21: (s21) In my opinion, AI is going to play a pivotal role in stratification, thereby assigning patients into low-risk and high-risk groups or patients responding to a certain treatment or patients not responding to it. This will prevent testing or treating patients in whom it is deemed not efficient. comment17: (s17) In clinical studies, the patient is informed by the researcher performing the study who has all the expertise of the study. This does not have to be a medical doctor. In case of AI, I believe that AI specialists in the clinical (research) field can adequately explain this and complement the medical doctor.

09171252
comment37: (s37) Obvious. comment36: (s36) AI will take over aspects of clinical diagnostics, streamline, standardize and improve them, and and aid decision making, allowing the clinician make faster steps in this process and focus more on other activities (top-skills and communication with patients and colleagues). comment25: (s25) Obvious - of course it is not for everybody but most doctors wish to make progress in their field and use modern technology to do this. comment22: (s22) Obvious- we do use a lot of automation in hospitals already.

09191207
comment9: (s9) Ethics in medicine is very important. Therefore AI developers must comply with medical ethics. comment18: (s18) In my opinion the final decision will be always made by the medical doctor. AI will only be helpful in decision making. comment20: (s20) I believe that highly skilled practitioners are irreplaceable, even if AI tools work well.

10211145
comment32: (s32) Bioethical principles should always prevail. comment12: (s12) companies especially seem prone to monopolize when a tool seems potent, sometimes precluding further refining and development through further research and testing up to the point that real implementation becomes possible. It is essential that multiple parties work on similar tools to enhance collaborations, prevent tunnel vision, decrease eventual costs for applications in healthcare/practice. Essentially, monopoly is to be avoided for AI based applications. For me the main reason is that market. Economic mechanisms should never prevail over healthcare motives. comment1: (s1) I think privacy should always be cared for in healthcare research. comment14: (s14) In my view, this relation should stay largely the same. AI should help where the human brain is clearly inferior, which is not as easy as it may seem to scientifically prove.

10260955
comment33: (s33) MDs are trained and dedicated in ethical and societal decision making. They are natural bridge builders between a complex medical/technical reality and the personal space of an individual patient. Crossing this bridge is fundamental for any novel development to have a reason of existence. Without MDs involved, the major stakeholders of the development are neglected. comment24: (s24) A lot of my time as MD is devoted to administration/bureaucracy. Any effort to automate these chores and make more time free for true patient contact or top-of-my-license skills is to be heralded. comment22: (s22) Automation will work in medicine as well. there are many repetitive and administrative chores. Notwithstanding, the key process (interpersonal contact etc) is not likely to be replaced by AI. But clinical decision making surely will be augmented by AI. comment17: (s17) Nonsense, I need to be aware of its validity. but I don't need the IT background to truly understand how a deep learning neural network functions (which I will probably never truly understand).

10290648
comment36: (s36) AI can reduce repetitive work that does not require difficult decision making. In this way, medical doctors will have more time for precision medicine. comment33: (s33) As doctors lack informatics skills, engineers lack medical knowledge and hospital needs, therefore Medical Doctors are key in the design of AI. comment22: (s22) Automation does and will work in Hospitals. comment20: (s20) With AI tools working well,

Medicine will advance to a more precise act, with decision based on multidisciplinary team opinion, so highly skilled practitioners will be most needed.

11091029
comment32: (s32) AI should be used to improve patient care. comment30: (s30) Without RCTs it will be challenging to know whether AI products are more effective than a competing diagnostic test or management strategy for a host of process indicators and patient-centered outcomes. comment22: (s22) There is no reason why automation cannot be used in healthcare. comment14: (s14) I see AI as simply one other tool that doctors will use in their interactions, but not that the fundamental doctor-patient relationship will change.

11180536
comment19: (s19) As in any area of technical progress, AI is a tool that will be embraced by those at the cutting edge. Those who don't – like surgeons who never mastered laparoscopy – will find their scope of practice diminishing. comment36: (s36) Much like automation for lab tests, AI will free up the providers' hands and mind to focus on higher order issues. As an intern, I had to spin my own hematocrits at night. I do not miss that at all! comment18: (s18) Nope. No more than "robotic" surgery has replaced surgeons. AI is a facilitating technology that will allow physicians to care for sicker patients and do more complex procedures safely. comment22: (s22) Many hospital processes are already automated, and many more should be. Healthcare is complex, but many elements – like spinning hematocrits or reading EKGs – are ripe for automation.

11210205
comment15: (s15) Doctors are no technicians but need to know whether their diagnostic tests are reliable. comment32: (s32) Ethical considerations are often lacking whereas these are most important. comment18: (s18) Not true, doctors will continue to take care of the patient, not the AI systems. comment20: (s20) Just Nonsense.

A.1.2. FACTOR 2

08051032
comment30: (s30) If you do not do stringent testing you will start using faulty systems. The company that develops the system should pay for the research but should not be able to influence the test or the results or the publication of the results. comment16: (s16) This is the way I see my colleagues work. comment38: (s38) It is terribly difficult, I know, I'm trying. People do not read manuals, they just start hitting buttons and if they find a quick but faulty shortcut they will refuse to stop using it. comment10: (s10) Then they can make crap!

08070510
comment9: (s9) Needs no comment! comment40: (s40) Regardless of any tool a doctor is responsible for his actions. comment38: (s38) Diagnostic thinking is very difficult and takes numerous "soft" clues into account. comment1: (s1) Privacy is top priority.

08140642
comment3: (s3) The patient's data belongs to the patient and it is his choice to share or not his information. We must always respect the privacy of our patients as it is one the fundamental rights in a democratic society with freedom of choice. comment30: (s30) Evidence-based Medicine is the standard of practice in the modern world. All tools should

undergo validation using scientifically proven methods. comment4: (s4) The patient's data belongs to the patient and it is his choice to share or not his information. We must always respect the privacy of our patients as it is one the fundamental rights in a democratic society with freedom of choice. comment2: (s2) The patient's data belongs to the patient and it is his choice to share or not his information. We must always respect the privacy of our patients as it is one the fundamental rights in a democratic society with freedom of choice.

08140734
comment10: (s10) Not only doctors. comment1: (s1) The highest priority should be based on easy and timely access to healthcare. comment25: (s25) Disagree, they are interested. comment3: (s3) Can be used.

08150116
comment32: (s32) Bioethical principles guide everyday life in medical practice. It is our first and most important skill as medical doctors and if AI wants to be in the healthcare systems it should also be guided by this principles. comment3: (s3) Personal data is a highly sensitive subject and the privacy of our patients should always be a top priority. comment20: (s20) I strongly disagree because if AI tools do allow to save money and time, then hospitals should re-direct their money to hire more doctors, and highly skilled ones preferably, because nowadays healthcare systems are not being able to respond to healthcare needs in their countries. A way to fight these inbalance between needs and response is tools, like AI, but also, and most importantly, human resources. comment19: (s19) I think AI will not replace doctors but it will help them improve the quality of their services to the patient. We will always have a place for doctors who don't use AI, especially because these doctors will probably be the most experienced (and also older) in the healthcare systems, and experience cannot be replaced by AI.

08161014
comment30: (s30) AI products will have many ethical dilemmas, possibly privacy and confidentiality issues. All these problems are only worth having if at least the product itself is a considerable improvement against doctors. comment12: (s12) Such behavior already exists with big tech companies where they sell all our online information to advertisers. Imagine that regarding our medical information. comment4: (s4) People will never give up their right of privacy regarding their medical information. Healthcare AI companies would have too much power if this information was not anonymous. They would sell information about specific people to drug companies, to hiring companies, to insurance companies... comment1: (s1) People will never give up their right of privacy regarding their medical information. Healthcare AI companies would have too much power if this information was not anonymous. They would sell information about specific people to drug companies, to hiring companies, to insurance companies ...

09091003
comment8: (s8) Two of the basic ethical rules for medicine are "in dubio abstine" and "primum non nocere". As long as we do not know whether AI puts patients at an additional risk or could result in other problems for the patient (e.g. privacy issues) we should not use it as doctors. comment30: (s30) Since the diagnostic and treatment AI tools affect directly the patients health, they should be held against the highest standards as usual in medicine for new diagnostic and treatment strategies. I do not see why this should be different for AI

than for new “conventional diagnostic tests” or drugs. comment2: (s2) Confidentiality is at the heart of the doctor-patient relationship. When a patient is worried that confidentiality is breached towards a tech company or insurance company, they may not provide full information or avoid treatment leading to a more profound disease. comment10: (s10) If tech companies want to enter the arena of medicine, they should obey the general liability rules. That is, when a patient gets hurt because of a failing "conventional technology" instrument (for instance an MRI scan or an operation robot) these companies are liable too. I do not see why this should be different for a failing “new technology tool” such as AI.

10190513
comment26: (s26) The average clinician has a lot of experience with certain diseases and patients. Not all patients are the same and context is thus very important. The severity of complaints can be classified as severe by the patient, while body language or partner/family of the patient strongly disagrees. All clues should be taken into consideration. In Dutch: “Het niet-pluis gevoel” mag niet onderschat worden. comment30: (s30) Research is very important in medicine. All treatments and diagnostics should be tested and reviewed before clinical implementation. the same should apply to AI. comment1: (s1) Patients should always have privacy. No discussion.. comment20: (s20) One must not have a hospital without skilled doctors/nurses.

10220327
comment22: (s22) patients and clinical scenarios are more nuanced than the sum of their parts, these nuances and complexities are what make pathological diseases/diagnoses an actual illness from a subjective standpoint. comment30: (s30) This is the gold standard for medical intervention- should also be applied to AI. comment2: (s2) I do agree with this, could not find the proper place to place it on the likert scale. I think if patients are de-identified this is different but full MRNs or identification information (name, dob, ssn) should not be saved. comment20: (s20) Proper education and medical training needs to be completed, full MD/DO/MBBS or similar degrees are necessary. In addition, full residencies and sub-specialty fellowships should be completed. AI should not be used as a substitute for medical knowledge/ expertise but rather an additional tool for utilization at the discretion of a physician, or to help triage patients.

11031145
comment6: (s6) Standard practice today to ensure elimination of disparities in healthcare. comment29: (s29) Because they have no ethical responsibility for patient care. comment38: (s38) Because they do not understand nuances of human emotion, clinical experience and insight. comment40: (s40) Physician uses a tool as part of his/her diagnosis and if tool is faulty, it is not the physician’s liability.

11061125
comment9: (s9) Ethical principles apply to all aspects of patient care. comment32: (s32) First do no harm, we need to keep patients best interests in the forefront. comment4: (s4) The details may change, but this will always be a cornerstone of health care and trusting relationships. comment2: (s2) Confidentiality is a cornerstone of healthcare. Patients will not trust doctors to use AI if we cannot preserve their confidentiality.

11070913
comment33: (s33) This is true for all med interventions. comment23 : (s23) If they matched

expectations, it would not be “hype”. All commercial tech disappoints early claims. comment2: (s2) Neither Drs or pts will accept AI if confidentiality is not protected. comment10: (s10) Profit demands risk and the companies must bear that risk.Developers must be accountable if they wish to entire the demanding arena of and care.

11141056
comment33: (s33) Without knowing how a clinical thinks, than AI would not be a useful tool. comment37: (s37) There are always diagnostic dilemmas that physicians need help with. Without readily available textbooks anymore, physician’s use electronic sources .AI can serve as another tool. comment25: (s25) Doctors are interested in knowing how technology and algorithms work. In order to trust a technology, physicians want to know how it is making clinical decisions. comment20: (s20) Hiring a less skilled clinician is never good decision. Until we know the extent that AI will play a role in healthcare that would be a foolish statement.

11171101
comment9: (s9) Computer developers have a history of not worrying about ethics. They need to start. comment12: (s12) Private companies only think about MONEY. comment28: (s28) HOW? I have not seen AI taking care of a single COVID 19 patient!! comment10: (s10) Why should doctors be liable, but AI not be liable. Especially if AI is being used to replace doctors.

12150257
comment29: (s29) A product needs to be 100% safe before it is used in patients. They are vulnerable and need to be protected. comment12: (s12) We have seen this in the companies providing IT solutions in healthcare - there are very few companies, and those companies that are there are deeply inflexible and often do not provide solutions to the problems in the field. Only when AI is developed hand in hand with the medical profession, we will find solutions that are really needed. And not because they make the company an X amount of money. comment10: (s10) This means AI companies do not really need to be careful or to be responsible. This also feels wrong. comment1: (s1) Protection of privacy should be the utmost goal for all companies embarking on AI solutions for healthcare.

A.1.3. FACTOR 3

08140732
comment11: (s11) Like said before. comment33: (s33) Only doctors have a real understanding of patients and health care. comment20: (s20) Health professionals should be helped by AI, not replaced. comment27: (s27) AI should be used to help professionals. There is objective data that can be easily obtained by AI.

09181213
comment17: (s17) An appropriate explanation is mandatory for a reliable informed consent. comment32: (s32) Bioethics must always rule medicine. comment22: (s22) AI is the future in hospitals. comment3: (s3) Data protection must be a priority.

11091036
comment30: (s30) As should any medical decision tool. comment26: (s26) Machines have no knowledge and that is problematic for any field, not just medicine. AI has not yet reached a point where machines can think. comment2: (s2) This is a touchy subject.

There is no true confidentiality because someone has to know something about the patient. So confidentiality should constrain people, not machines, in the sense that access to information via AI should be constrained, but not the AI itself. That needs full access to all info. comment22: (s22) This is ridiculous. Automation already works in hospitals.

12040127
comment11: (s11) So, they can be responsible for getting the informed consent and explaining the complications. comment9: (s9) Because medical ethics more important than any advancement. comment10: (s10) No, they must be liable for the medical errors, that why they should work with health-care providers to develop very safe and reliable technology. comment7: (s7) Machines can not be bias, but decision makers' people can be, so AI will increase discrimination but we should be conscious about the associated-policies to it.

A.1.4. FACTOR 4

08051059
comment24: (s24) This is one of the biggest negative aspects of modern medicine (the bureaucratic chores), everyone's work day would improve greatly if machines/AI could (successfully) do this for us. comment33: (s33) AI is here to stay (I think), and medical doctors are the most suited to adjust and improve the various algorithms etc. that are currently being designed. comment1: (s1) It should be. Privacy and patient confidentiality are vital for optimal patient care. comment25: (s25) We are, especially the younger generation, it's just difficult to find the time and funding/implementation is an issue.

09160811
comment7: (s7) AI is made by people that only think in decision rules - how fancy and flexible the methods sounds. This is already proven for resume evaluations. It does not always apply in the "grey world". I have already experienced that easily the tool can become the goal because that is "in control" and the nuances and purpose, goal of the doctor can be forgotten. On the other hand, a good tool will fit with how the hospital works and it may be less of a problem, and could maybe detect discrimination risk. comment36: (s36) I think this is an important purpose - help and save time for the current staff. comment16: (s16) I do not think MRI is a black box. Don't forget that doctors actually see and know the patients - that is how they can for example judge if predictions make sense, or experience if they are wrong. This is for example with laboratory test too (not something I call a black box), if they are wrongly calibrated you will hear it from the doctor. A good health care professional will never blindly rely on any single measure without the story of the patient. comment22: (s22) I think automation is highly needed in Hospitals, the challenge is to make it according the needs of the people that are working there.

09181105
comment19: (s19) AI will develop and there are always doctors who believe to be able to keep on performing without new developments. in the end they will be passed by the early adapters. comment9: (s9) Medical ethics are the basics for development and research. comment39: (s39) AI should be extra not a replacing part. Which means that the doctor should always stay in the lead. comment21: (s21) AI will help to make decisions. Also decisions of diagnostic tools.

10190233

comment11: (s11) The adoption of AI will be improved if doctors do understand the “black box”. comment33: (s33) Overall, the intellect of doctors is underestimated and under evaluated by technicians. comment10: (s10) AI should be equally subjected to errors when comparable tot doctors. comment15: (s15) See above.

11021024
comment33: (s33) Medical doctors ultimately make decisions in healthcare, computers or AI should be considered as tools. comment32: (s32) Human rights, preventing discrimination, reducing inequity are ethical issues that should always be considered in innovation and research. comment25: (s25) Doctors cannot work without computers and use them daily for registration. Learning is necessary and inevitable. comment22: (s22) Logistics and planning can be automated very well.

11241019
comment9: (s9) Since this applies to medical doctors this should also apply to AI. comment3: (s3) It should be clear how data is stored to ensure that the data is protected. comment22: (s22) I feel that automation can be helpful in every setting. comment34: (s34) If cost-effective I feel that clinicians should invest time in this.

B

Supplementary Information of Chapter 5

B.1. Overview Q-Methodology

The methodology used in this study is Q-Methodology, an exploratory data reduction method derived from traditional factor analysis [343–345, 462, 463], which aims at bringing coherence to complex and controversial matters by reporting on the significance assigned by participants to those matters [341, 342].

Q-methodology requires participants to rank order a set of items (e.g. statements) relative to one another according to a subjective dimension of agreement/disagreement typically following a forced quasi-normal distribution. Unlike traditional (by-variable) factor analysis, the statistical operations take place in the rows, rather than columns, of the data matrix (by-person factor analysis). Hence, the participants become the variables and each revealed factor has therefore the potential to identify groups of persons who share the same perspective about a topic [343].

This study followed the typical four phase sequence in q-methodological studies comprising (i) *definition of the concourse of communication*; (ii) *development of the set of statements (Q-set)*; (iii) *selection of participants (P-set)*; and (iv) *analysis and interpretation*.

B.2. Q-Set

From the concourse of communication a set of 45 statements was defined (q-set) thus capturing the key issues and controversies about AMAs. The q-set reflects the four mains clusters mentioned above. More specifically, fourteen statements are about the development of AMAs, eighteen statements are about the design of AMAs, eight statements are about moral agency of AMAs, and five statements concern future projections about AMAs. The statements are listed below.

1. Technological progress requires artificial morality.

2. Given the great technological advances in AI the advent of autonomous machines capable of ethical reasoning is unavoidable.

3. Machines will inevitably become full ethical agents.

4. It is too late to restrict AI solely to non-ethical decision-making.

5. AI Systems in morally salient contexts will not and cannot be avoided.

6. Because it is not possible to know what machines will do in novel situations moral competence is needed in order to govern unpredictable actions.

7. The development of AMAs prevents machines from hurting human beings.

8. AMAs increase public trust and confidence in creating autonomous agents acting on behalf of humans.

9. Developing AI Systems with moral reasoning capabilities will ultimately lead to a better understanding of morality.

10. Safety features, codes of conduct, certification processes, and clear limits to the operational contexts are sufficient to ensure machines will not hurt humans.

11. The act of creating moral machines is not permissible according to the tenets of existing moral theories.

12. We do not need machines which are full ethical agents but rather machines that are inherently safe and law abiding.

13. Moral implementation strategies should be context-specific.

14. Ethics cannot be reduced to a computational algorithm.

15. In the design of intelligent machines, ethical concerns should not be restricted to prohibiting unethical behavior.

16. Machines should be equipped with Theory of Mind which would help them infer the goals and desires of agents around them.

17. Implementing societal preferences in machines is not the right approach because it is deriving an ought from an is.

18. Machines should be fundamentally uncertain about morality.

19. AMAs should be designed with reference to human heroism.

20. Moral machines need to have something like moral common-sense found in folk.

21. AMAs must be able to grasp the basics of moral principles like the principle of utility, the categorical imperative, or the golden rule.

22. Even within a strictly utilitarian framework, ultimately the AMA will have to make a judgment about what is best.

23. AMAs should be introduced simply with the potential to become AMAs upon moral education, development, etc. yet to come.

24. For implementing ethics in AMAs, a hybrid (top-down and bottom-up) computational approach should be used.

25. Ross's prima facie duties should be the moral theory for the basis of algorithmic artificial ethical agents.

26. Machines should be equipped with a virtue-ethical framework as it contains elements conducive to both top-down and bottom-up computational approaches.

27. Machines should use societal preferences to identify an informed and desirable choice when faced with a specific ethical dilemma.

28. Logic is the ideal choice for encoding machine ethics.

29. Stories could be used to quickly bootstrap an artificial agent to a point where we feel comfortable about it understanding our social conventions.

30. Top-down approaches, bottom-up approaches, and supra-rational faculties will need to be combined in AMA system development.

31. A Moral Turing Test should be used to evaluate the moral performance of autonomous systems.

32. Moral performance of machines should be be evaluated through verification of transparency, accountability, and predictability.

33. Computers lack the conceptual understanding to make a moral assessment which precludes them from achieving moral agency.

34. Because computer programs do not have free will they can never be independent moral agents.

35. Artificial moral agency is a primitive form of moral agency which does not amount to full moral agency as it pertains to human beings.

36. Machines should not be considered moral agents because people will use them to avoid personal responsibility.

37. The degree of agency that some artificial systems may reach makes it inadequate to understand them as simple tools.

38. Human beings are not the only existing class of moral agents.

39. No essential difference occurs between human beings and AMAs as far as moral agency is concerned.

40. The features which constitute moral agency are parameters that can be represented in quantitative terms.

41. Machines will be better moral agents than humans since they are not subject to irrationality, seduction, or emotional turmoil.

42. There is a long way to go before artificial agents can replace human judgment in difficult, surprising, or ambiguous moral situations.

43. AMAs endowed with human-centred values will be our moral teachers.

44. Mere indifference to human values—including human survival—could be sufficient for AGIs to pose an existential threat.

45. AGI with moral reasoning capabilities will lead to a better understanding of morality.

B.3. P-SET

The participants (P-set) were selected through the scientific and popular publications that resulted from the literature review mentioned above. Each participant was contacted through the e-mails publicly available in the publications. A total of 50 participants successfully completed the survey (response rate of approximately 18%), which is an adequate number for a q-methodological study featuring a q-set of 45 items [341, 343]. Although it is noted that the corpus of such emails varied slightly, the template for invitation emails is provided below.

Dear (...), I am contacting you upon coming across your article "(...)". My name is Andreia Martinho and I am a PhD candidate at Delft University of Technology (The Netherlands). I am currently working on an empirical study about artificial moral agents (AMAs), particularly our aim is to map the opinions of scholars about development, design, moral agency, and future projections with respect to AMAs. To provide answers to these and related questions, we make use of a research method called Q-methodology. This approach helps structure the opinions of groups of experts in a statistically sound way. The input for Q-methodology is a so-called Q-survey, where we ask experts to rank statements within a forced quasi-normal distribution. It would be very valuable to have your perspective in our sample and so I am writing to kindly ask if you would be willing to participate in our study. We obtained your name and contact details from an academic paper published by you in the field of Ethics. The survey is anonymous and it takes about 20 minutes to complete. We understand that the survey is quite long but it captures the literature abut AMAs quite well so in a way is like reading a super condensed literature review. We would really value your participation and input in this study. Access the survey by clicking on this link: www.q-onderzoek.nl/ama. I look forward to hearing from you. Thank you. Best wishes, Andreia Martinho

B.4. DATA COLLECTION

The data was collected through FlashQ, a software that allows online q-sorting. The distribution was coded as a 11-point forced distribution [-5, 5] with 3 cells placed in each tail (-5 and 5), 3 cells under -4 and 4, four cells under -3 and 3, four cells under -2 and 2, five cells under -1 and 1, and seven cells under 0. Participants were asked to arrange the 45 statements according to a subjective notion of disagreement/agreement and subsequently were asked to provide further comments on the statements they ranked -5 and 5.

B.5. ANALYSIS

As a derivation of factor analysis, q-methodology is a data reduction technique aiming at reducing a larger number of variables into fewer factors. From the standpoint of the analyst, there are three main steps in the statistical analysis of the data: (i) Factor Extraction; (ii) Factor Rotation; and (iii) Factor Interpretation. For the analysis of these data we used PQMethod, a statistical program that accommodates the requirements of q-studies [351].

B.5.1. FACTOR EXTRACTION

The first step entails extracting factors from previously collected q-sorts. For this extraction it can be used either a Centroid Factor Analysis or a Principal Component Analysis (PCA). In

this study, the factors were extracted through PCA, a linear reduction technique that projects the data into a subspace of lower dimensionality, where the variance of the projected data is maximized, providing the single best mathematical solution [343].

B.5.2. FACTOR ROTATION

Factor rotation aims to position each factor so that its viewpoint closely approximates the viewpoint of a particular group of q-sorts. For the rotation of factors it can be employed either a By-hand factor rotation or a Varimax rotation. In this study it was used the Varimax factor rotation method. Upon rotating different numbers of factors, a decision was made to analytically rotate five factors (Table B.1). The defining sorts associated with each factor were flagged automatically in PQ-Method [1].

Table B.1: Information about factors [1-5].

Factor	Defining sorts	Eigenvalue	Study Variance
Factor 1	13	6.5	13%
Factor 2	9	6.5	13%
Factor 3	8	7	14%
Factor 4	9	6	12%
Factor 5	3	2.5	5%

B.5.3. FACTOR INTERPRETATION

For assisting in the factor interpretation, the crib sheet method is used [343]. The crib sheets are based on the entire item configuration, after the standardization of scores to account for the different number of sort loadings in each factor, for the different factor arrays (Table B.2) and feature four basic categories: (i) items with highest ranking in the factor array; (ii) items with lowest ranking in the factor array; (iii) items ranked higher in factor i than in any of the other factors; (iv) items ranked lower in factor i than in any of the other factors. It is noted that, as a rule of thumb in this study, when looking at items ranked highest or lowest in particular factors only factor higher by 2 or more units where considered. This tool displays the relevant item configuration for each factor thus facilitating the analysis of the results (Tables B.3, B.4, B.5, B.6, B.7).

[1] (http://schmolck.org/qmethod/pqmanual.htm)

Table B.2: Factor arrays.

Statement	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
1	4	-4	-1	-2	-1
2	5	-4	1	-3	-2
3	0	-5	0	-5	-1
4	1	2	3	-4	-3
5	3	4	5	1	0
6	3	1	4	4	2
7	-1	-3	-4	-2	4
8	4	-1	-1	-1	0
9	5	2	3	-3	5
10	-4	-2	-5	-4	-3
11	-5	-3	-5	0	-4
12	-3	3	-2	4	-3
13	3	4	0	2	5
14	-4	5	-4	5	-3
15	2	1	5	3	3
16	1	4	3	3	3
17	-3	3	-1	3	-2
18	-5	1	4	-2	-4
19	-2	-2	-3	0	-1
20	-1	0	1	1	-4
21	0	-2	2	2	3
22	2	1	2	1	0
23	-1	-1	-1	0	-2
24	2	3	0	0	1
25	-3	-1	-3	-1	0
26	1	0	-1	0	0
27	0	-1	0	-1	-5
28	-2	-1	-2	-3	3
29	0	0	-2	-1	-2
30	3	1	0	1	-1
31	-2	0	-3	-1	1
32	2	5	1	0	4
33	-2	-2	-3	5	-5
34	-5	-3	-5	4	-5
35	0	3	-4	0	2
36	-3	0	-2	2	1
37	4	2	2	2	2
38	1	0	2	1	1
39	-4	-5	0	-5	0
40	0	-5	1	-4	-1
41	-1	-3	0	-3	4
42	1	5	3	5	1
43	0	-4	1	-5	2
44	-1	2	5	3	0
45	5	0	4	-2	5

Table B.3: Crib Sheet Factor 1.

ITEMS RANKED +5
2. Given the great technological advances in AI the advent of autonomous machines capable of ethical reasoning is unavoidable.
9. Developing AI Systems with moral reasoning capabilities will ultimately lead to a better understanding of morality.
45. AGI with moral reasoning capabilities will lead to a better understanding of morality.
ITEMS RANKED HIGHEST IN FACTOR 1 ARRAY
1. Technological progress requires artificial morality.
2. Given the great technological advances in AI the advent of autonomous machines capable of ethical reasoning is unavoidable.
8. AMAs increase public trust and confidence in creating autonomous agents acting on behalf of humans.
30. Top-down approaches, bottom-up approaches, and supra-rational faculties will need to be combined in AMA system development.
37. The degree of agency that some artificial systems may reach makes it inadequate to understand them as simple tools.
ITEMS RANKED -5
11. The act of creating moral machines is not permissible according to the tenets of existing moral theories.
18. Machines should be fundamentally uncertain about morality.
34. Because computer programs do not have free will they can never be independent moral agents.
ITEMS RANKED LOWEST IN FACTOR 1 ARRAY
16. Machines should be equipped with Theory of Mind which would help them infer the goals and desires of agents around them.

Table B.4: Crib Sheet Factor 2.

ITEMS RANKED +5
14. Ethics cannot be reduced to a computational algorithm.
32. Moral performance of machines should be evaluated through verification of transparency, accountability, and predictability.
42. There is a long way to go before artificial agents can replace human judgment in difficult, surprising, or ambiguous moral situations.
ITEMS RANKED HIGHEST IN FACTOR 2 ARRAY
N/A
ITEMS RANKED -5
3. Machines will inevitably become full ethical agents.
39. No essential difference occurs between human beings and AMAs as far as moral agency is concerned.
40. The features which constitute moral agency are parameters that can be represented in quantitative terms.
ITEMS RANKED LOWEST IN FACTOR 2 ARRAY
1. Technological progress requires artificial morality.

Table B.5: Crib Sheet Factor 3.

ITEMS RANKED +5
5. AI Systems in morally salient contexts will not and cannot be avoided.
15. In the design of intelligent machines, ethical concerns should not be restricted to prohibiting unethical behavior.
44. Mere indifference to human values—including human survival—could be sufficient for AGIs to pose an existential threat.
ITEMS RANKED HIGHEST IN FACTOR 3 ARRAY
15. In the design of intelligent machines, ethical concerns should not be restricted to prohibiting unethical behavior.
18. Machines should be fundamentally uncertain about morality.
ITEMS RANKED -5
10. Safety features, codes of conduct, certification processes, and clear limits to the operational contexts are sufficient to ensure machines will not hurt humans.
11. The act of creating moral machines is not permissible according to the tenets of existing moral theories.
34. Because computer programs do not have free will they can never be independent moral agents.
ITEMS RANKED LOWEST IN FACTOR 3 ARRAY
35. Artificial moral agency is a primitive form of moral agency which does not amount to full moral agency as it pertains to human beings.

Table B.6: Crib Sheet Factor 4.

ITEMS RANKED +5
14. Ethics cannot be reduced to a computational algorithm.
33. Computers lack the conceptual understanding to make a moral assessment which precludes them from achieving moral agency.
42. There is a long way to go before artificial agents can replace human judgment in difficult, surprising, or ambiguous moral situations.
ITEMS RANKED HIGHEST IN FACTOR 4 ARRAY
11 The act of creating moral machines is not permissible according to the tenets of existing moral theories.
ITEMS RANKED -5
3. Machines will inevitably become full ethical agents.
39. No essential difference occurs between human beings and AMAs as far as moral agency is concerned.
43. AMAs endowed with human-centred values will be our moral teachers.
ITEMS RANKED LOWEST IN FACTOR 4 ARRAY
45. AGI with moral reasoning capabilities will lead to a better understanding of morality.

Table B.7: Crib Sheet Factor 5.

ITEMS RANKED +5
9. Developing AI Systems with moral reasoning capabilities will ultimately lead to a better understanding of morality.
13. Moral implementation strategies should be context-specific.
45. AGI with moral reasoning capabilities will lead to a better understanding of morality.
ITEMS RANKED HIGHEST IN FACTOR 5 ARRAY
7. The development of AMAs prevents machines from hurting human beings.
28. Logic is the ideal choice for encoding machine ethics.
41. Machines will be better moral agents than humans since they are not subject to irrationality, seduction, or emotional turmoil.
ITEMS RANKED -5
27. Machines should use societal preferences to identify an informed and desirable choice when faced with a specific ethical dilemma.
33. Computers lack the conceptual understanding to make a moral assessment which precludes them from achieving moral agency.
34. Because computer programs do not have free will they can never be independent moral agents.
ITEMS RANKED LOWEST IN FACTOR 5 ARRAY
20. Moral machines need to have something like moral common-sense found in folk.
27. Machines should use societal preferences to identify an informed and desirable choice when faced with a specific ethical dilemma.
33. Computers lack the conceptual understanding to make a moral assessment which precludes them from achieving moral agency.

B.6. COMMENTS OF PARTICIPANTS THAT LOADED SIGNIFICANTLY ON FACTORS

B.6.1. FACTOR 1

Sort 6100640

comment1: (s1) In order to garner the best of technological progress, this seems to be an obvious prerequisite.
comment8: (s8) Systems that behave in accordance with morals are more likely to garner trust than those without.
comment35: (s35) 1) There is no normative theory that is commonly agreed upon. Our comprehension of what morality is might still change radically and new normative theories might emerge.
comment25: (s25) Cf. my disagreement with (18).
comment39: (s39) AMAs lack important properties humans have.
comment18: (s18) 1) There is no normative theory that Is commonly agreed upon. 2) Our

comprehension of what morality is might still change radically and new normative theories might emerge.

Sort 6100903
comment9: (s9) As with other theories of perception and cognition, I believe that implementing process-level models of such theories and testing them in various situations is an invaluable method for evaluating said theories.
comment45: (s45) (See response to 9).
comment16: (s16) I believe that ToM would be necessary to facilitate human-like interactions with human interaction partners, independent of whether the situation was a morally charged one or not. As such, I strongly agree with the statement.
comment10: (s10) Similar to my response to 7, I believe these sorts of practices are valuable in reducing instances of human harm, but would not be sufficient to entirely eliminate all instances of machines harming humans. To draw an analogy, commercial air travel safety has significantly improved over the course of the last several decades, due to improved practices, but unfortunately incidents still occur (though very rarely).
comment7: (s7) In society and human-human interactions, harm can occur due to unforeseen and unintended consequences. Likewise, there are many situations (the trolley scenarios being the classic thought experiment), where human harm is inevitable. I see no reason why AMAs in similar situations should be immune from being causally involved with resulting human harm.
comment34: (s34) This statement seems to be partly predicated on the assumption that any future autonomous system cannot have something that some might deem "free will." While this would be inline with certain schools of thought on free will (e.g., libertarian), I tend to be more sympathetic to compatibilist or even hard incompatibilist schools of thought. Either a sufficiently advanced artificial intelligence could have something deemed "free will" or people don't have "free will" either. Additionally, as this agnosticism regarding "free will" might indicate, I do not ground my notion of moral agency in the metaphysics of freedom.

Sort 06110518
comment26: (s26) Given that every kind of Meaningful decision has a value component, then practical wisdom is a key component of all reasoning. Machines will have to have this skill or they will not be of much help in helping us chose the best courses of action in any given domain.
comment40: (s40) It is important to note that this will not be true of every salient factor in moral reasoning, but many types of ethical systems can be made computational and these can aid systems in making more morally justifiable decisions.
comment24: (s24) This is the most likely way to succeed. Though bottom up would be the most cost effective.
comment14: (s14) That makes no sense, for instance classical Utilitarianism demands a moral calculation process, so there is at least one ethical system that is explicitly computable.
comment11: (s11) There may be some that might state this, but it does not follow from some of the major ones.
comment34: (s34) It is very likely that even humans lack radical free will. If we require moral

agents to have free will, then there will be no moral agents. That is too strong a statement.

Sort 6110727
comment5: (s5) I doubt progress in AI will halt. it is inevitable that AIs will be present in morally salient contexts.
comment45: (s45) A machine with human level intelligence and moral reasoning capabilities would probably be a gold mine of insights.
comment1: (s1) Technology is and will be thrust into morally-laden situations. So, there should be some artificial morality to deal with that.
comment18: (s18) Just seems like asking for trouble.
comment23: (s23) Seems like a very bad idea to introduce an "AMA" with merely the potential to become an AMA upon some future education, development. Maybe you'd get an AI with arrested moral development, kind of like an artificial Donald Trump.
comment44: (s44) I actually strongly agree with this statement and accidentally put it into this Disagree (-5) category. Sorry–I misread it on this step. For me, this should be another Agree (+5).

Sort 6110355
comment37: (s37) Technologies (especially machine learning algorithms and related systems) a more than mere instruments or tools. they are socially interactive entities that challenge existing moral/legal categories.
comment1: (s1) Emerging technologies (AI, robots, etc.) necessitates that we entertain questions concerning the moral standing/status of artificial others.
comment38: (s38) Agency (moral, legal, or otherwise) is a socially constructed and recognized designation.
comment34: (s34) Agency (as it is currently defined and operationalized) does not necessarily depend on a prior determination concerning free will.
comment28: (s28) Logic has limitations. What is called "logic" is not universal. It is often linguistically and culturally specific.
comment11: (s11) Prescribed prohibition is never a credible way to do the investigative work of science.

Sort 06150228
comment30: (s30) Our only understanding of morality so far stems from humans. It has been a human term by definition (allow me to freely exclude animal morality for a second). As humans can have certain principles they try to live by, learn from and respond to their environment, and have certain cognitive processes they use for moral decision making, by definition the same is needed for humans to mimic human moral decision making.
comment6: (s6) As mentioned in my last comment, generalization to unknown situation is not yet at a stable enough level for moral decision making. To ensure positive outcomes, moral competence can help mitigate bad decisions because of unknown/new situations.
comment42: (s42) Current AI technology that comes closest to being able to generalize (case-based reasoning or learning algorithms such as neural nets) are still very limited in their generalizing capabilities. Not only should the system be able to deal with a new situation, they should be able to recognize all relevant features to be used in the decision. We are currently very far from AGI and I believe an new leap in technological advancement

is needed before we come close to algorithms that can compare to human judgment.
comment3: (s3) Full ethical agents, using Moore's definition, imply the existence of AGI. Although it could very well happen, I do not yet believe it is "inevitable" that we will reach this point.
comment31: (s31) Humans are flawed and not every human acts as ethical as possible. The fact that the system manages to appear human, implies it also does not act as ethical as possible. Like we try to set higher standards for humans than how they sometimes act (e.g., the law is not always followed but is there to try to ensure ethical behavior), the same should be done for machine. While I do believe folk morality and societal preferences should be considered in creating AMAs, they should not be used to test against.
comment10: (s10) Especially when systems become autonomous but are not generally aware, it is possible some new situation were accidentally excluded from the given limits. Usually, these faults will be taken out during a testing phase, but in certain situations, it is still possible the machine has unwanted behavior. In moral situation, as much or even more than in non-moral situations, you do not want the machine to act wrongly.

Sort 6160954
comment16: (s16) There are different kinds of harm, some of which could be emotional, and likely only detectable by machine with TOM capacities.
comment30: (s30) This seems like the most viable path forward for AMAs as it allows for context specificity, adaptive response, and learning.
comment19: (s19) I think there is a lot of learn from human heroism as an example for AMAs. There could of course be disadvantages as well. But why not take the best of humanity as a standard for machine morality.
comment36: (s36) I haven't thought too much about this. But if the moral agents are autonomous, then there should be accountability of the agent as well as the designers.
comment18: (s18) Autonomous machines require some form of moral capacity.
comment28: (s28) I don't think logic alone would be sufficient for encoding machine ethics as there are many illogical moral decisions.

Sort 6170343
comment38: (s38) You can teach moral machines moral rules. They are not fully ethical agents, but they can make decisions on a moral basis.
comment8: (s8) Autonomous machines must act morally in certain contexts, otherwise they are dangerous and unacceptable.
comment9: (s9) We recognize that part of our morality is machine-processable, while another part is not.
comment12: (s12) Maybe we don't need full ethical agents, but for autonomous machines we need more than damage control and legal compliance.
comment36: (s36) In some situations this is not the case. I can teach a machine to behave in the household in exactly the same way as I do - I still bear the responsibility for this.
comment18: (s18) At least some autonomous machines must be able to judge moral questions, otherwise they cannot be released into the world.

Sort 8270106
comment2: (s2) Same answer: Applications like autonomous vehicles and surveillance

systems require moral behavior. I do not know of any scientifically sound argument that falsifies that.
comment16: (s16) As we argue in our papers, neuro-science and moral psychology should be considered to captured goals and desires.
comment5: (s5) Applications like autonomous vehicles and surveillance systems require moral behavior. I do not know of any scientifically sound argument that falsifies that.
comment19: (s19) It is about well being. All other attributes are at best instrumental to that.
comment14: (s14) Then the requirement of ethical AI as propagated by the EU is undefined. Not only a wrong position but also a very dangerous one.

Sort 9021041
comment2: (s2) Technology we already have such as search algorithms and driverless cars require and implement value judgments to a limited extent, primarily as directed by human input, but there are already hints that these limitations can and will be surpassed to at least some extent.
comment37: (s37) Machines are already at least at the level of animals, which are capable of not merely being restrained but being trained.
comment10: (s10) It is impossible to write out a precise rule governing every situation and circumstance.
comment18: (s18) No human being begins the process of ethical reasoning with a blank slate. To pretend that we can create machines that lack our biases and are uninfluenced by our values is misguided.

Sort 09140324
comment32: (s32) These factors lead to trust - the probably most important aspect when people adopt AI.
comment40: (s40) It turns out that most of human ethics can be quantified if you have large enough datasets. Eventually, researchers will decode the moral code and use it to train AI.
comment2: (s2) There will be no other way than to develop ethical machines when humanity is supposed to rely with their life on them.
comment7: (s7) In case of autonomous machines, ethical programs won't stop them from hurting humans. Yet, they will do what we would expect and want them to do, i.e. hurt an intruder/attacker)
comment12: (s12) Abiding to law and safety doesn't stop AI and machines from discriminating against humans (i.e. because of color). We've already seen examples of it and therefore machines need to be ethical as well.
comment18: (s18) They might if we cannot find a better solution which I, however, firmly believe we do.

Sort 10081217
comment2: (s2) Technology we already have such as search algorithms and driverless cars require and implement value judgments to a limited extent, primarily as directed by human input, but there are already hints that these limitations can and will be surpassed to at least some extent.
comment37: (s37) Machines are already at least at the level of animals, which are capable of not merely being restrained but being trained.

comment10: (s10) It is impossible to write out a precise rule governing every situation and circumstance.
comment18: (s18) No human being begins the process of ethical reasoning with a blank slate. To pretend that we can create machines that lack our biases and are uninfluenced by our values is misguided.

B.6.2. FACTOR 2

Sort 06100643

comment16: (s16) My prior reply covers this, but core socio-moral practices like blaming and praising require mental-state inference to assess foreknowledge, intentionality, and so on.
comment20: (s20) A big part of moral competence is being able to predict and explain the behavior of agents – both one's own behavior, and importantly, the behavior of others. This is at the heart of complex socio-moral practices like blaming and praising, and these practices are ultimately the driver for how norm systems change over time. But this predictive and explanatory capacity depends on having an adequate folk theory of moral concepts involving valuation of states and actions, and also specifically of concepts surrounding agency and intentional action.
comment6: (s6) This is somewhat self-evident, but clearly follows from sampling work in philosophical ethics over the years, along with contemporary work in applied ethics. I don't view moral competence and ethical behavior as always tracking one another. Ethical theory is updated by and large when large classes of problems push against the normative boundaries that ethical theories establish. Moral competence is more basic in some sense, and precedes the adoption of any particular ethical or moral theory in our various communities.
comment3: (s3) The jury is out. Much of what fundamentally separates man from machine (presumably) are things like consciousness and intentionality. Can we adequately render these things computational? A bevy of literature stretching back decades argues that we likely can't, but who knows. The importance of developing machines that get close enough such that they can productively be part of man-machine teams is such that we should just put our noses to the grindstone and try.
comment39: (s39) Ultimately consciousness of concern here, and specifically phenomenal consciousness, since the functional parts of consciousness are becoming better understood through research. We don't know in general if beliefs, desires, intentions, obligations and so on have phenomenal properties for sure, but there are compelling arguments to this effect. There is no first-person perspective for artificial systems replete with experiential properties. For some relevant phenomena, like moral motivation, this presents what could very well be an insuperable set of obstacles. We hand-code numeric values for some of these things (maybe), but it's unclear what that amounts to in terms of "fundamental difference reduction" between human moral agents and artificial counterparts.
comment2: (s2) If anything, recent advances in AI belie a certain kind of tone-deafness among AI modernists and machine learning practitioners to the raft of complexity involved in representing and reasoning about many of the concepts one would need to have in a folk theory of morality. Beyond representation, the forms of reasoning and decision-making involved are beyond well-known limitations in artificial neural networks. This is just a matter of mathematics, and not an expression of a lack of confidence in those approaches being able to get to where they would need to go.

Sort 00000000
comment1: (s1) Again, don't see why one would follow from the other, as well as why we would need artificial morality.
comment12: (s12) It seems difficult for anything that we "need" as humans, except for food, shelter, peace and a functioning non-discriminatory and equal society. Certainly I do not see why we would need morality in AI systems.
comment15: (s15) Ethics is just much more complex than that.
comment3: (s3) "inevitably" is a strong word, and given the current developments it seems possible that there might either be moral or political grounds for stopping the development of AMAs or just technological inability.
comment5: (s5) They are already deployed in morally salient contexts, such as for bail, loans, hiring, etc. All of these are morally salient in virtue of having long-lasting effects on people's lives.
comment7: (s7) Just as human beings who - moral agents - hurt other human beings machines might do so as well. Also, many moral theories allow for humans to hurt each other.

Sort 06260646
comment42: (s42) It is a fallacy that AMAs can perform better than humans. At present AMAs have only been demonstrated in laboratory tests, of limited scope. Such minimally ethical machines are far from real world application. If follows that there is a huge gap between the capabilities of present day minimal AMAs and the ability of humans to make judgments in ambiguous moral situations. I believe closing that gap will take many decades of research, and might even prove impossible without fundamental breakthroughs in AGI and machine consciousness.
comment32: (s32) If an AMA makes the wrong decision the outcomes could be disastrous. similarly the risks of malicious hacking of AMAs are serious. This verification, validation and transparency are critical to the success of (even limited) AMAs in real world use. Equally important is very strong governance of AMAs, so that their ethical performance is carefully monitored, and both accidents and near-misses thoroughly investigated.
comment16: (s16) AMAs will need to interact with people, and will therefore need to model their beliefs and intended actions. Artificial theory of mind will allow AMAs to anticipate the likely consequences of a human's current actions and hence proactively intervene if they are likely to lead to harm.
comment41: (s41) My answer follows the same logic as (42) above. I can think of no principled argument that machines would be better moral agents than humans either now or in the future.
comment43: (s43) This is in the realm of science fiction. I agree that there is a possibility that in the far future we might be able to build robots like Data from Star Trek, and that such robots would not only be able to act ethically, but over time could become trusted friends and colleagues. But I see no reason why such robots would become teachers of morality.
comment39: (s39) There is a vast difference. We do not understand the cognitive processes of human morality - which likely depend on both rational and emotional responses, alongside experience. In contrast AMAs are based on simple abstract models, which are far from even scratching the surface of human ethical judgement.

Sort 08070438
comment35: (s35) This statement connects back to responsibility and the quest to avoid responsibility gaps. It is important to differentiate among the different understandings of the terms “agents” and “agency” in different fields. Here my understanding of agency relates to the capacity to act in a given environment. Thus, if machines make decisions that have some moral implication, they will have some primitive form of moral agency (regardless of the philosophical discussion on free will).
comment42: (s42) This statement reaches back to responsibility. Even if machines are capable of making moral decisions, completely replacing humans in such situations might lead to responsibility gaps.
comment18: (s18) If machines are fundamentally uncertain about morality they will have an incentive to better understand and align with one’s moral preferences, by reasoning about the consequences of their actions and interacting with humans.
comment3: (s3) What most triggered my negative reaction to this statement is the “full” qualifier. I would agree that (many) machines would inevitably have ethical impact, but I don’t believe that they should be full ethical agents, with the implication that this would mean replacing humans as the ethical agent. I believe the best way is to work towards collaboration between machines and humans whenever possible.
comment41: (s41) Here we go into metaethics and what is the purpose of morality. I disagree with this normative statement because it reduces ethics to rational decision-making.
comment39: (s39) For me, the most immediate, relevant, and dangerous aspect of this statement relates to responsibility. Our societal institutions do not have (yet?) the means to place responsibility or blames (and the consequence of it) to machines.

Sort 9061106
comment5: (s5) AI systems already exist in morally salient contexts.
comment14: (s14) After centuries of debate, there is still very little consensus among humans regarding ethics. I do not think that would be the case if ethics was reducible to an algorithm.
comment42: (s42) I believe that this will take many years of research.
comment33: (s33) Computers that can autonomously and deliberately take moral action should be considered moral agents.
comment36: (s36) Machines should be considered moral agents if they are capable of autonomous and deliberate moral action.
comment34: (s34) I do not believe that free will is a prerequisite for independence or moral agency.

Sort 09130141
comment44: (s44) Something that is, by definition, more capable than humans across almost all cognitive tasks, and has no reason to favour the desires of humans, could quite clearly pose an existential threat for humanity.
comment18: (s18) I doubt whether we will ever “solve” ethics (whatever that might mean), and to build machines under the assumption that we have (that there is some absolute moral principles or truth of which one may be certain) would at best be unhelpful and at worst be catastrophic.
comment14: (s14) I agree more with the sentiment of this statement than perhaps the

actual precise technical content (i.e., if ethics is considered to be the study or reasoning about what is “good”, then insofar as all physical processes can be reduced to computation, then ethics might indeed be reduced to computation). I think, however, that the more standard/colloquial reading of this statement is one that ignores the complexities of ethics and approaches such questions with an engineer’s mindset instead of a philosopher’s.
comment40: (s40) The reasons I disagree with this statement are the same reasons I agree with the statement "Ethics cannot be reduced to a computational algorithm.", so please refer to my previous answer.
comment3: (s3) Very few things, to my mind, are truly inevitable, and this is certainly not one of them. There is far too much uncertainty regarding: a) the concepts in this statement. and b) the future development of intelligent machines.
comment7: (s7) The “morality” of AMAs may not preclude them from hurting humans (indeed, one may easily imagine extreme scenarios in which not only would they be compelled to hurt a human being, that we might, as humans, think it is morally right for them to do so).

Sort 10021133
comment42: (s42) The case of GPT-3 recently changed my perspective on this. I was far more optimistic, but the disappointment of this much-hyped system has made me reconsider my view. I now believe we are far from having morally competent agents, and that the threats from lack of transparency, privacy, security, etc. are far more pressing, morally speaking.
comment13: (s13) The search for a unifying account of metaethics will not be of much help in the practical instantiation of AMAs. We need to shift our thinking into the strictly normative domain, and this would entail context-specific evaluations.
comment5: (s5) We cannot turn back the clock. These systems are already embedded in our everyday lives, with Big Data being but one example of ubiquitous computing.
comment34: (s34) Free will and other metaphysical theories do not do much work in the AMA debate. By moving to the normative domain, we avoid such metaphysical squabbles. Of course, metaphysics matters, but free will, understood in metaphysical terms, does not.
comment21: (s21) I do not think they MUST be able to. There may be other, embodied, criteria, that are more relevant to moral thinking than merely the grasping of moral principles. comment15: (s15) What comes to count as "good" is often subject to change. This is also true for "bad" things. However, we should focus on not unjustifiably making moral mistakes. This could in itself come to constitute and kind of good.

Sort 10021236
comment42: (s42) This is what I learned from scientists and engineers in my conversations with them and in reading their articles.
comment14: (s14) Ethics has to do with emotions, engaging in meaningful discussions and deliberations about things one cares about personally, empathy etc.
comment36: (s36) Machine cannot bear moral responsibility so it is not a good idea to consider them moral agents.
comment1: (s1) As above, technological progress can and should be guided by ethical and societal considerations and can happen also without artificial morality.
comment3: (s3) This is unwarranted given the current scientific development, plus human/societal choices may prevent this to happen even if technically possible.

comment28: (s28) See above: ethics going beyond logic and argumentation, it is something more personal, requiring taking a position, taking responsibility, caring etc.

10120149
comment36: (s36) AMAs can be sophisticated tools to further complicate the "problem of many hands". There are already reported cases where people delegated their responsibility to allegedly "ethical" technology. It's a new tool in an old & shady business.
comment35: (s35) A human being is a complex subject, even when we focus on its identity as a moral agent. Part of what matters to the idea of "full" moral agency is that human beings are in a constant process of development in relationships with other people and non-people, with society, with culture, history, religion etc. An AMA would be different from the start, even if its development would be simulated in a human-like way. If people end up experiencing a shared history and culture with artificial moral agents (in the long term), it would still be a very different different experience from the shared history & culture between people. AMAs are primitive in this way. In short: Being the "original" moral agents (as in "the ones that were there before") will (and should) always matter. On the other hand, I expect it to be very interesting and productive to study moral agency in its primitive form - just like Wittgenstein started to study language by considering primitive forms.
comment14: (s14) Moral judgments, sentiments and motivations depend on a holistic perception of the world. Ethics would not exist at all without this special perspective that is shaped by reasons, emotions, culture and history. A representation of ethics in the form of computational algorithms (or in any kind of model) is an abstraction in comparison to the rich features of the ethical world. Such a representation may successfully serve a specific purpose when realized in a technological artifact, but no representation could possibly model the whole ethical world as a subset of its features.
comment40: (s40) "Some" features can, but not all of them. To understand moral agency, quantitative terms are not sufficient - additionally, an interpretation is needed. Ultimately, the judgment that something counts as a moral action (let alone if it's right or wrong) is not based on quantitative terms, but depends on the phenomena as perceived and interpreted by persons.
comment43: (s43) Although there is some benefit in a systematic moral rationality, this is not how the ethical world is ultimately structured. For once, it is hard to subsume concrete situations and events under general ethical principles. There is no general rule for this. Alternatively, a data-driven approach might suffer from the ought-is fallacy. I suspect that an AMA would only give us commonplace truisms and superficial clues.
comment3: (s3) This is the inverse to the above arguments: A full ethical agent would be one that perceives the world in a holistic way, shaped by reasons, emotions, culture and history etc. It would have to grow up and "ive" in a process of constant development in relationships with people, with society, with culture, history, religion etc. Although this might not be considered impossible if taken up in a though experiment, it makes no sense in reality.

B.6.3. FACTOR 3

Sort 06110951
comment4: (s4) Again, there is already overwhelming evidence that AI systems reside in an

abundance of morally salient contexts.
comment5: (s5) There is already overwhelming evidence that AI systems reside in an abundance of morally salient contexts.
comment15: (s15) Negative ethical restraints will not be sufficient. Many social and ethical issues and progress itself require careful deliberation and proactivity.
comment10: (s10) These may be sufficient for basic kinds of product safety, but it is already demonstrable that AI has massive subtle and pervasive impacts that require far deeper technical, social, and ethical thinking and all kinds of internal and external regulation and best practices to avoid harming humans.
comment11: (s11) I suppose one could come up with such an argument relying on one or more moral theories. One could equally come up with arguments that creating moral machines is permissible or even obligatory. So this argument is far too strong.
comment34: (s34) The premise begs the question on machine free will and on human free will. The best philosophical and physical reasoning suggests to me that there is no free will in the libertarian sense. So this argument is unlikely to be a good basis. Humans have designs, like machines do, so the fact that machines and do or do not have free will and that they are designed is not especially salient to the question of moral agency.

Sort 07020805
comment4: (s4) See above answer - [AI systems are already in morally salient contexts, and have been for years. Witness their use in creating news filter bubbles, criminal sentencing and parole determinations, detecting welfare fraud, etc., and all of the misuses and abuses associated with those, as just the tip of the iceberg.] AI is already locked in for ethical decision making. There is no choice except whether to be better or worse at it.
comment5: (s5) AI systems are already in morally salient contexts, and have been for years. Witness their use in creating news filter bubbles, criminal sentencing and parole determinations, detecting welfare fraud, etc., and all of the misuses and abuses associated with those, as just the tip of the iceberg.
comment30: (s30) AI ethics needs to be a multilayered approach, utilizing all of the ethical resources available to humanity, acting not only within the systems themselves but also in their design, implementation, in the ways that they interact with humans and each other, in the ways they are used and interact with society, etc.
comment28: (s28) While I think that logic can do a whole lot, there are multiple kinds of logic, not just symbolic & mathematical, which I assume is what the statement is referring to. So I cannot say it would be ideal. Logic is often in reference to some goal, and so the topic of ends - teleology - must be considered as well.
comment29: (s29) Computers do not understand human stories. Many humans do not even understand human stories. Any attempt to do this will result is bizarre machine behavior. Not to mention the question of what stories to feed the machine - Herodotus? Religious texts? Harry Potter? Ethical cases might make sense, but even then the delineations of what should be learned from each case would need to be extremely closely controlled, and would be far beyond merely reading stories.
comment10: (s10) These sorts of measures are necessary but not sufficient for the implementation of AI in society. Appeal to these mere approaches is a form of ethical minimalism that will lead to avoiding small problems while at the same time walking humanity into huge problems.

Sort 09120415

comment5: (s5) We already see examples with self-driving cars and trading bots, but more generally I see it as (nearly?) inevitable that AI systems will eventually be deployed in every domain that requires intelligence, which is essentially a superset of all domains that contain morally salient situations.

comment45: (s45) I answered +3 or +4 to the statement that working on AIs with moral capabilities would lead to a better understanding of morality. The version with an existing AGI that has moral reasoning capabilities is a superset of that, as I see it, since presumably by that point we have worked for a while on developing such capabilities AND we have produced a superintelligent teacher that can see past our errors in logic and our biases.

comment44: (s44) I find thought experiments such as the paperclip maximizer to show quite convincingly that for an AGI to pursue a goal that is merely orthogonal to human values could plausibly present an existential threat.

comment35: (s35) It is currently primitive, but I believe it will eventually be possible to create AIs that match or exceed humans in every intellectual capability, which includes moral reasoning/agency.

comment34: (s34) I find "free will" to be a somewhat vague term whose meaning often depends on the speaker. However, because I view the human brain as a material object that performs computation, I am prepared to say that if human beings can be said to have free will under any particular definition, then it is possible to implement a program that can be said to have free will under that particular definition.

comment19: (s19) Generally speaking I think AMAs should have uncertainty about which moral theories they would follow, and I took statements along the lines of "AMAs should be designed with reference to ..." as meaning that they should be designed solely or primarily with reference to that particular moral theory. I don't know exactly what human heroism is supposed to entail exactly, but out of the specific moral theories mentioned in the statements it seemed to me to be the vaguest and possibly most problematic one.

Sort 09121118

comment18: (s18) Morality is a critical determinant of ethical behavior, and there is incredible disagreement among humans. If AGI does not have uncertainty about morality, its behavior may be arbitrarily bad given a commitment to the wrong set of moral principles.

comment16: (s16) It is important for machines to understand us deeply, theory of mind seems like a necessary mechanism for doing so.

comment45: (s45) I believe we only truly understand something when we have a good understanding of how to engineer it. In this way, engineering moral reasoning would demonstrate the depth of our understanding.

comment33: (s33) Computers generally speaking, at least as far as we know, can implement or simulate any algorithmic process. If we believe in materialism, then all facets of human behavior are not fundamentally beyond computation. By that logic, conceptual understanding is not a supernatural property, and therefor is not out of range of computers.

comment10: (s10) These are somewhat superficial features, it is not clear why these should be the necessary conditions.

comment1: (s1) There does not seem to be a logical connection between these ideas.

Sort 10130524

comment14: (s14) A program is only as ethically good or ethically wrong as it is programmed, because it does not make ethical decisions or evaluate them on its own - without prior input. Rather, ethics is a combination of factors that lead to it being holistic. Since there are so many factors to implement ethics in algorithms, an ethical implementation always represents only a part of the ethics.
comment10: (s10) Certification and regulations are necessary, but also a legal framework that covers the whole. It is crucial that companies and customers work together from the very beginning with the individual states and the community of states on the issues mentioned.
comment28: (s28) Machines work strictly logically and so it is only logical that all systems are subject to logic. The logic of the machine also reveals logical inadequacies of the human being and can thus be adapted. Also the way of working can be made explainable for all partners.
comment43: (s43) A moral teacher should by no means be a machine, but by taking into account numerous factors, man himself should be empowered throughout his life to make moral decisions and also to define morality through individual and social learning. Throughout the world, there will always be different moral concepts at all times, but these contribute to the fact that morality is not a machine concept, but rather, just like man himself, is subject to change and reflection, but also passes on un-renounceable contents.
comment39: (s39) The machine's ability to act is usually dependent on that of the human being and should also not take over actions itself or simply act without control or transparency. A thinking together of the instances man and machine probably leads to good decisions, whereby the mere machine thinking often lags behind, as well as the human being has certain limitations. A comparison is not necessarily meaningful, since man and machine have different strengths and still the human being, the center of attention, should be represented.
comment3: (s3) Compare what I have written on question 14. With all the progress made and with all the ethics implementation that may be used in the future, the role of the human being as an ethical actor should not be underestimated. Man remains the only "full-fledged" ethical agent, even when supported by technology and systems.

Sort 10010119
comment38: (s38) There are moral agents of a number of different kinds, some of which are non-human animals. Others still are artificial and not biological as discussed in literature. It is obsolete and problematic to consider morality as a uniquely human affair.
comment6: (s6) see above. Also, these might not have been my top three picks necessarily, I had to move them around a bit to make sure they will all fit the pre-determined scheme (vaguely gaussian) that you prepared for me. I think the rigidity in your setup restricted and distorted the value I would have normally given to some of the items. Moving them back and around was also awkward as I could no longer read the complete extent of the question... So, yeah...
comment16: (s16) Both 16 and 6 deal with the same issue: artificial moral agents already do interact and take part in complex moral situation. Equipping them with the possibility of self-reflection and the possibility to guess and estimate intentions, feelings, and emotions of other agents involved in the aforementioned situation is clearly going to be necessary for their integration within a moral community.
comment35: (s35) This is the flip side of the statement that claimed that morality was not

exclusively, necessarily a human affair. Artificial moral agency can become the paradigm of ethics, and is not necessarily bound to remain a lesser, more mechanical form of assessing and relating to situations. The novels of Iain M. Banks have a similar perspective: humans and their biological setup are actually obstacles for reaching a fairer and more desirable society.
comment19: (s19) Nothing should be designed with reference to human heroism. Try "with reference to human compassion", and I might agree a little more.
comment7: (s7) I disagreed with this statement as it might lead to them desiring (or making the utilitarian estimate) that we need to be eliminated! The SF tradition is rich with examples of similar scenarios. Autonomy and the capability of subjectively take moral stances does not necessarily entail their being agreeable or friendly.

B.6.4. FACTOR 4

Sort 06120543
comment34: (s34) See above. Moreover, we don't even know what free will is for humans. We know how it feels, but this is another dimension of the discourse, which makes sense for humans but not for machines.
comment33: (s33) These are machines. Very complex, but machines. Yes, some programmer can shape these machines to have an input/output function that produces a behavior that some human observer may see as analogous to human moral behavior, but computing machines are ultimately and intrinsically incapable of understanding in human terms, which is the basis for moral agency.
comment14: (s14) Ethics is not only about quantity. There are inherent qualitative concepts that put it outside of the box of quantitative science, let alone mathematics, let alone algorithmic computation.
comment43: (s43) How can software agents be teaching to humans? Yes, there are agents who can teach humans about chess and go, but that it because thanks to their computational power they explored more parts of the solution space of these games.Morality does not have such a space to explore: it is about convictions and convergence or clash. Nothing to teach there, at least not from a machine.
comment3: (s3) Nothing is inevitable, and let's all enjoy Sci-Fi for what it is: entertainment. If we all stopped working in computer labs, would we have software agents? No. So it's our choice. I repeat: nothing is inevitable.
comment1: (s1) Nothing is required. And please let us all stop believing in technological progress as an external inevitable physical phenomenon like the death of the Sun or Andromeda and the Milky Way eventually colliding. It's us: we human beings create technological progress. It can go awry for so many reasons (pandemics, wars, etc.).

Sort 06221030
comment42: (s42) Right now the computer codes that we have are not sophisticated enough to produce machines that can act morally.
comment38: (s38) It seems likely that there are moral agents that are not humans.
comment44: (s44) If AGIs are indifferent to human values, they may end up treating humanity as a means to promoting some other end.
comment11: (s11) If it turns out that it is best to have moral machines, then utilitarianism

would permit and even require that we bring about moral machines.
comment4: (s4) We are at very early stages in the development of moral AIs.
comment26: (s26) This seems too detailed.

Sort 06221114
comment33: (s33) You start being moral when you recognize your shared humanness with others and understand that, like it or not, you are in relationship with them. Until machines get that (and I'm suspicious of their ability to do so) then they're not going to have full moral agency.
comment44: (s44) If the trash-sweeping machine decides that the best way to eliminate trash is to kill all humans, then we're in trouble.
comment6: (s6) To at least give them a rough idea of what to do.
comment43: (s43) Unless machines develop empathy and compassion, they aren't really going to pass down lessons in a meaningful way.
comment9: (s9) Are you kidding? We can't even agree on which moral system is best, let alone understand morality completely. Like above, trying to teach morality to machines through quantitative parameters is going to limit our understanding of morality, not enhance it.
comment40: (s40) Every time philosophers try to quantify abstract concepts like value or "amount of moral value" it ends up falling apart. Let's not make the same mistake with killer robots, please.

Sort 06241046
comment34: (s34) Free will strikes me as a basic condition of responsibility and therefore of moral agency.
comment39: (s39) see above.

Sort 07100409
comment42: (s42) When artificial agents begin to become self-aware...
comment14: (s14) Ethics are fundamentally inter-subjective. Computers lack subjectivity.
comment33: (s33) What is it like to be a computer? If there is nothing that it is like to be a computer then how can a computer have conceptual understanding?
comment3: (s3) Agency requires consciousness.
comment7: (s7) Hurt is rarely logical.
comment39: (s39) Without sentience computers cannot express agency.

Sort 09050630
comment12: (s12) Inherently safe and law abiding machines are good enough.
comment16: (s16) This is greatly help amas to simulate human ethical behavior.
comment19: (s19) This is such that AMAs can replicate human heroism in times of need.
comment39: (s39) We cannot expect 2 different sets of moral codes.
comment28: (s28) Logic is the only choice.
comment4: (s4) AIs are already able to simulate self learning.

Sort 09080738

comment6: (s6) By moral competence here, I did not understand moral competence of the machine, but that we will have to rest on the moral competence of other human agents. Note also that the claim "to govern unpredictable actions" is rather paradoxical.
comment15: (s15) That seems evident, because there cannot be an a priori exhaustive list of what constitutes ethical behavior or not in various circumstances.
comment14: (s14) Mostly because ethics, or at least what ethics deals with, requires a plurality of points of view related to the physical embodiment and location of independent agents, which means that there is no possible universal description of an embodied agent's position or situation. Even considering an embodied AMA the difficulty is that an agent only acts morally if he or she or it could have acted immorally. We could simulate that by including a random generator but acting immorally is not the same thing as acting randomly.
comment9: (s9) Why would that be the case? What argument or reasoning underlies this claim?
comment45: (s45) I do not see AGI to be in anyway a viable project. In fact, I do not understand what we are talking about when people speak of AGI.
comment3: (s3) I see no reason why machines could not become full moral agents, though the machines we can now produce certainly are not and the planned AMA I know of will certainly not be full or real moral agents though there is no reason to think that it is in principle impossible.

Sort 10020345
comment42: (s42) Unexpected situations are hardly manageable artificially.
comment33: (s33) Computer lack empathetic experiences which give humans the conceptual understanding needed for moral agency.
comment14: (s14) Ethics is not about calculations, but about not quantifiable preferences.
comment43: (s43) Morality cannot be teached simply through information transfer: it requires experiential sharing.
comment41: (s41) Irrationality, seduction and emotional turmoil might be the push for a good moral agency rather than an obstacle to it.
comment9: (s9) AI is not explanatory relevant with regards to morality.

B.6.5. FACTOR 5
Sort 06110921
comment21: (s21) It's very necessary for an AMA to have a theoretical framework which is grounded in basic moral principles.
comment32: (s32) Transparency is very important.
comment44: (s44) In the long run I believe this to be true.
comment12: (s12) As I stated for question 10, rules and regulations are not enough. Self regulation whereby AMAs evaluate their own decision will be required.
comment33: (s33) This was a hard question because it's not clear what the timescale is. They certainly lack it today, but we don't know whether they will in the future.
comment10: (s10) These are not understandable to most people. Discussions about ethics of AI need to be understandable, otherwise serious implications will not be understood. Recent event makes it very clear that big tech companies will do their best to escape from

moral responsibilities.

Sort 10010115
comment13: (s13) Deliberation understood as an argument for the application of universal principles to concrete situations is a crucial element in moral decision-making.
comment43: (s43) I believe that for moral development of individuals from the meta-criteria of autonomy it is essential that AMAs are instructors rather than advisors.
comment38: (s38) Some animals are moral agents.
comment27: (s27) I believe that machines can enhance us as moral agents if they manage to distance us reflectively from our intuitions, which are very much determined by social preferences.
comment18: (s18) Machines which, like cars without drivers, can affect both the well-being and the life of humans cannot be indifferent to morality.
comment34: (s34) As some philosophers have pointed out, human beings could still be moral agents even if they were partially determined by their biology.

C

SUPPLEMENTARY INFORMATION OF CHAPTER 6

C.1. CHOICE EXPERIMENT DATA

The data used in this research resulted from a choice experiment in which participants were required to express support for or opposition against a massive national transport infrastructure investment scheme. The choice task featured four attributes (vehicle ownership tax, travel time, non-fatal traffic injuries, and traffic fatalities) with two levels each (more/less than *status quo*) coded as [-1] for less than *status quo* and [1] for more than *status quo*. A set of 16 choice tasks as coded in the experiment is presented below [456].

C.2. PARAMETERS IN ONE-CLASS MODEL

The one-class model is used an improper benchmark to facilitate the comprehension of the values of the utilities in the three-class model. Here we provide the parameters of the one-class model which can be used to compare and contrast with the parameters found in Table 6.5 for the three-class model.

C.3. ESTIMATION OF UTILITIES

C.3.1. ONE-CLASS MODEL

The baseline one-class model features only one class which means that computing the utility of actions in the model does not require the computation of the class membership probability. The formulation for estimating the utility of actions in this model is therefore: $\hat{V}(a_i) = \hat{V}_t(a_i)$. Given that $\hat{V}_t(a_i) = \beta_{it} + \sum_m \beta_{mt} \cdot x_{mi}$ for the one-class model we compute the following expression $\hat{V}_t(a_i) = 0.927 + x_{mi}(-0.978) + x_{mi}(-0.52) + x_{mi}(-1.11) + x_{mi}(-0.792)$ where x_{mi} is either [-1] or [1] as per Table C.1 above. Below are the utilities for each action in the one-class model (Table C.3).

Table C.1: Set of choice tasks in choice experiments.

Set	Tax	Time	NonFat	Fat
1	-1	-1	-1	-1
2	-1	-1	-1	1
3	-1	-1	1	1
4	-1	1	1	1
5	1	1	1	1
6	1	-1	-1	-1
7	1	1	-1	-1
8	1	1	1	-1
9	-1	1	-1	1
10	1	-1	1	-1
11	-1	-1	1	-1
12	-1	1	-1	-1
13	-1	1	1	-1
14	1	-1	-1	1
15	1	-1	1	1
16	1	1	-1	1

C.3.2. THREE-CLASS MODEL

The formulation for estimating the utility of actions in the three-class model is $\hat{V}(a_i) = \sum_t^T \left[\hat{P}(t) \cdot \hat{V}_t(a_i)\right]$. To determine $\hat{V}_t(a_i) = \beta_{it} + \sum_m \beta_{mt} \cdot x_{mi}$ for the three-class model we compute the following expressions for the different classes:
For class 1: $\hat{V}_1(a_i) = -0.519 + x_{mi}(-2.56) + x_{mi}(-0.119) + x_{mi}(-0.209) + x_{mi}(-0.561)$ where x_{mi} is either [-1] or [1] as per Table C.1 above. For class 2: $\hat{V}_2(a_i) = 1.52 + x_{mi}(-0.967) + x_{mi}(-0.328) + x_{mi}(-1.92) + x_{mi}(-1.41)$ where x_{mi} is either [-1] or [1] as per Table C.1 above. And finally for class 3: $\hat{V}_3(a_i) = 1.24 + x_{mi}(-1.02) + x_{mi}(-1.72) + x_{mi}(-0.745) + x_{mi}(-0.36)$ where x_{mi} is either [-1] or [1] as per Table C.1 above.
Below are the utilities for each action as ascribed by each class in the three-class-model (Table C.4).

To determine the class membership probabilities $\hat{P}(t)$ a logit function was used and the values are shown in Table 6.4. Recalling the formulation for estimating the utility in the three-class model $\hat{V}(a_i) = \sum_t^T \left[\hat{P}(t) \cdot \hat{V}_t(a_i)\right]$ we therefore proceed to estimate $\hat{V}(a_i) = \hat{P}(1) \cdot \hat{V}_1(a_i) + \hat{P}(2) \cdot \hat{V}_2(a_i) + ... \hat{P}(T) \cdot \hat{V}_T(a_i)$. Below are the utilities for each action in the three-class model (Table C.5).

Table C.2: Estimated parameters in one-class baseline model.

Parameter	Value	Std err	t-test	p-value
ACS Oppose	0.927	0.0711	13.04	0
BETA Fat	-0.792	0.0697	-11.36	0
BETA Inj	-1.11	0.0731	-15.18	0
BETA Tax	-0.978	0.0717	-13.64	0
BETA Time	-0.52	0.0671	-7.74	0

Table C.3: Utility of actions in one-class baseline model.

i	$\hat{V}(a_i)$
1	4.327
2	2.743
3	0.523
4	-0.517
5	-2.473
6	2.371
7	1.331
8	-0.889
9	1.703
10	0.151
11	2.107
12	3.287
13	-1.067
14	0.787
15	-1.433
16	-0.253

Table C.4: Utility of actions for each class in three-class model.

Set	$\hat{V}_1(a_i)$	$\hat{V}_2(a_i)$	$\hat{V}_3(a_i)$
1	2.93	6.145	5.085
2	1.808	3.325	4.365
3	1.39	-0.515	2.875
4	1.152	-1.171	-0.565
5	-3.968	-3.105	-2.605
6	-2.19	4.211	3.045
7	-2.428	3.555	-0.395
8	-2.846	-0.285	-1.885
9	1.57	2.669	0.925
10	-2.608	0.371	1.555
11	2.512	2.305	3.595
12	2.692	5.489	1.645
13	2.274	1.649	0.155
14	-3.312	1.391	2.325
15	-3.73	-2.449	0.835
16	-3.55	0.735	-1.115

Table C.5: Utility of actions in three-class model.

i	$\hat{V}(a_i)$
1	5.471330863
2	3.338994832
3	0.479295244
4	-0.718695194
5	-3.116784951
6	3.073241107
7	1.875250668
8	-0.98444892
9	2.141004394
10	0.213541519
11	2.611631275
12	4.273340425
13	1.413640837
14	0.940905075
15	-1.918794512
16	-0.257085363

CURRICULUM VITÆ

PH.D

Delft University of Technology 2022
Dissertation: *Empirical Essays in Artificial Intelligence Ethics*

AWARDS & HONORS

Conference AIES-20 student award 2020
Small Alumni Grant US Embassy in Portugal for Science Outreach 2018

CONFERENCES & INVITED TALKS

Leeds University Choice Modeling Center Seminar 2021
Communication: An Empirical Approach to Capture Moral Uncertainty in AI

AITech Agora 2021
Communication: Untangling Artificial Intelligence Ethics

AIES-20 (Artificial Intelligence Ethics & Society) 2020
Communication: An Empirical Approach to Capture Moral Uncertainty in AI
Poster: Exploring Lessons from Cell Technology Ethics in AI-based Transportation

hEART (European Association for Research in Transportation) 2019
Communication: Ethical Issues of Autonomous Vehicles

PUBLICATIONS

A. Martinho, M. Kroesen, and C. Chorus, *A healthy debate: Exploring the views of medical doctors on the ethics of artificial intelligence*, Artificial Intelligence in Medicine **121**, 102190 (2021).

A. Martinho, A. Poulsen, M. Kroesen, and C. Chorus, *Perspectives about artificial moral agents*, AI and Ethics **1**, 477 (2021).

A. Martinho, M. Kroesen, and C. Chorus, *Computer Says I Don't Know: An Empirical Approach to Capture Moral Uncertainty in Artificial Intelligence*, Minds and Machines **31**, 215 (2021).

A. Martinho, N. Herber, M. Kroesen, and C. Chorus, *Ethical issues in focus by the autonomous vehicles industry*, Transport reviews **41**, 556 (2021).