# The Influence of Ethical Guidelines for AI on Developers of AI

Oetse Donner
Supervisor: Prof. dr. ir. Jan Broersen
Second reader: Dr. Baptist Liefooghe
Bachelor Kunstmatige Intelligentie
Universiteit Utrecht
7,5 ECTS
6-11-2020

**Artificial Intelligence is a technology that will influence the future. To initiate a more ethical approach of AI, ethical guidelines are created by various stakeholders. This research is aimed at finding out how these guidelines influence developers of AI. To gather data, a focus group was organized with seven students and one programmer who all had a connection to AI. The results from the focus group show different views on multiple categories that relate to the ethical guidelines for AI. A general consensus was displayed that legislation is needed before ethical guidelines will have a noteworthy influence. The discussion provides four proposals based on the results and the literature on how to start raising ethical awareness in the AI community before legislation is realized.**

# Table of Contents

# 1. Introduction

Artificial Intelligence (AI) systems are driving a revolution in computer science that will initiate global changes. These will affect politics, economy and society (Harari, 2017). AI technologies deployed in online services are already reaching billions of people. The global AI in social media market is estimated to grow from USD 0.6 billion in 2018 to USD 2.2 billion by 2023 (MarketsandMarkets, 2018). There are 3.6 billion people using social media worldwide (Clement, 2020) and machine learning software is proven to effectively influence people's decision-making (Matz, Kosinski, Nave, & Stillwell, 2017). To prevent undesirable events form happening and to guide this revolution, many ethical frameworks are emerging. These ethical frameworks provide guidance on topics such as, but not limited to, privacy, sustainability, non-maleficence and accountability. These are to be used when working with or developing AI. Attempts have been made to find similarities between these documents (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020; Greene, Hoffmann, & Stark, 2019; Jobin, Ienca, & Vayena, 2019; Zeng, Lu, & Huangfu, 2018). Others have criticized the guidelines on their effectiveness (Hagendorff, 2020; Mittelstadt, 2019). The amount of research on ethical guidelines for AI is increasing rapidly. This paper will investigate whether this growth is also influencing the people behind the technology, the developers of AI.

This will be done through a qualitative applied policy research that uses data which is collected during a focus group organized to gain knowledge about what developers of AI think of the ethical guidelines, and how they are influencing their work. In a previous quantitative research on the influence of an ethical code on programmers it was found that there was no significant effect on decision-making. This study did not go into detail as to why and urged for more research on what influences decision-making of computer scientists (McNamara, Smith, & Emerson, 2018). Although these results are valuable, it is important to continue on this road and look for explanations why they were not affected by the codes and how to improve the influence of ethical guidelines. Because this is a qualitative research of small proportion it is to be seen as an exploratory study, looking for interesting theories and topics for future research. This paper will focus on the question, **what is the influence of ethical frameworks, created to guide the ethical development of AI, on developers of AI according to developers of AI?** and the following sub questions: **What are ethical frameworks concerning ethical AI and what are their limitations? How do developers of AI engage with these frameworks? What could improve the influence of ethical frameworks?**

The following chapter will elaborate more on AI and its potential pitfalls. Then a theoretical background will be given to explain more about ethical guidelines and criticism raised in the literature, continuing with a method section on how the research was conducted and how the data was analysed. The section thereafter will contain the results, followed by the discussion, answering the research questions using the results and proposing how to strengthen the ethical base of AI. The paper will finish with a conclusion.

# 2. Context

## *Artificial intelligence*

Artificial intelligence has become an umbrella term for many technologies; therefore, it is hard to give an exact definition. The European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) has been working on a definition. They have published a document elaborating on the definition of AI and defined it as:

> Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.
> AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)
> (High-Level Expert Group on AI, 2019a, p. 1)

The AI HLEG gave another, similar definition in their document, *Ethics Guidelines for Trustworthy AI*:

> Artificial Intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.
> (High-Level Expert Group on AI, 2019b, p. 36)

This definition of AI will be held throughout the rest of the paper because it is more precise and informative.

The idea of AI has been around for about 65 years and research in AI is progressing every day. In 1955 the Dartmouth Conference proposal marked the birth of AI with the aim "... to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy, Minsky, Rochester, & Shannon, 2006, p. 12). Making AI a new field of research with the goal to break down human intelligence and simulate it in computer programs. Since then, AI has reached many milestones that were previously considered fantasies. Famous examples being the defeat of the chess champion Garry Kasparov by IBM's Deep Blue in 1997, the 2005 DARPA Grand Challenge where the first autonomous vehicles managed to drive across 100 kilometres in the Mojave Desert and IBM Watson's victory of the creative, language-based game Jeopardy! in 2011. These milestones are typical examples of encouragements for the AI-community, proving every time that they could *make* something that was able to do things that were thought to be impossible for non-humans.

### *Research priorities for robust and beneficial artificial intelligence*

60 Years after the Darthmouth Conference an open letter concerning the pitfalls of AI was published (Russell, et al., 2015). This letter was issued by the Future of Life Institute and has collected over 7000 signatures in support, including Stephen Hawking, Nick Bostrom and Stuart J. Russel. The letter states that the field of AI has made great progress over the last 20 years due to new methods, collaborative research and availability of data and processing power. Because of this progress, AI applications are being deployed and used in economically valuable technologies, changing society. The potential of AI is huge but is important to find a way to “reap its benefits while avoiding potential pitfalls" (Russell, et al., 2015, p. 3). That is, not only making more capable AI, but also maximizing its societal benefit. The letter urges expanded *interdisciplinary* research on making AI robust and beneficial for society.

So, according to leading AI experts, there is a problem with AI, and something needs to happen. The problem is that the potential of AI is unfathomable, it is not desirable to let everyone create AI without contemplating the greater societal and economic effects (Harari, 2017). In the media, the mentioning of malevolent use of AI is not uncommon. Project Mave, where Google helped the American Department of defence on how to use AI in for example drone surveillance systems (Cameron & Conger, 2018) or IBM helping the NYPD to produce a facial recognition system that could be used to search by skin colour (Joseph & Lipp, 2018) and the famous data breech of Cambridge Analytica, where unauthorized data was used to profile US voters and target them with personalized data (Lapowsky, 2019) are only three of many disturbing headlines that have been in the news. Even though the open letter is asking the AI community to be thoughtful about the power they might obtain, these examples all happened after the open letter was published.

Attached to the letter was a research priorities paper (Russel, Dewey, & Tegmark, 2015) on how to address the problem. In this proposal, three short-term priorities are mentioned. The first one being, *Optimizing AI’s Economic impact*, exploring both sides of the economic changes and how to cope with them. The second and third priorities are, *Law and ethics research* and *Computer Science Research for Robust AI.* Where the first two are mostly calling for a collaboration between economists, legal experts, political scientist, ethicists and computer scientists, the third priority is aimed directly at the developers of AI systems. It stresses the importance of autonomous systems to “robustly behave as intended” (Russel et al., 2015, p. 107) following with different areas where robustness research is desired. Due to the limited scope of this paper, it will focus on the second priority.

# 3. Theoretical background

### *Ethical Guidelines*

To prevent the unethical use and development of AI, tech-companies, governments and international institutions are creating universal principles for AI. In one report (Jobin et al., 2019), 84 papers containing some form of ethical guidelines for AI are analyzed in the search for a global agreement. Companies like Microsoft, IBM and Google are mentioned in the report alongside the Australian government and the European Union, showing the broad interest in the topic. More papers have emerged looking for overlapping themes

within the many different documents that concern the ethical development of AI (Fjeld et al., 2020; Hagendorff, 2020; Zeng et al., 2018). 88% Of the documents identified by Jobin et al. (2019) were released after 2016, showing the popularity of the topic. Also, they note that most of the documents were produced by private companies and governmental agencies, most of which come from economically developed countries. Another remarkable finding is that non-maleficence is mentioned significantly more often than beneficence. Guidelines are thus focusing more on preventing harm than promoting the benefits of AI (Jobin et al., 2019).

Another document on principled artificial intelligence that was published shortly after, giving an extensive analysis of thirty-six documents containing AI principles (Fjeld et al., 2020). Eight key themes were extracted and explained in detail using the common terms mentioned in the documents. The eight key themes found were privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility and the promotion of human values. They note that “the more recent documents tend to cover all eight of these themes", continuing to state that “these themes may represent the ‘normative core’ of a principle-based approach to AI ethics and governance". (Fjeld et al., 2020, p. 5) So even though there have been many documents emerging in a short time there seems to be a consensus on what the most important topics are surrounding ethical AI.

## *Concerns*

Even though the ethical frameworks seem to be reaching a consensus, it does not mean that the ethical frameworks are successful in reaching their goal. The precise goal of the guidelines is stated slightly different in most documents. Still, a common goal seems to be providing criteria for how AI ought to be “developed, deployed and governed” (Fjeld et al., 2020, p. 12). Although it seems like a noble cause, even this goal is being debated as some companies are being accused of ‘ethics washing’. This is, showing involvement in ethics to satisfy the public, but not actually putting it into practice (Johnson, 2019). Not only the goal, but also the means are criticized. There are papers that are positive about non-legislative sources, such as ethical guidelines or principles, but little research has been done about the influence of such codes on people who create computer programs (McNamara et al., 2018, p. 730). Brent Mittelstadt (2019) argues why a principled approach towards AI ethics differs fundamentally from a similar approach used in health. The arguments he gives have much in common with a paper published only short after by Thilo Hagendorff (2020) that also sheds light on the shortcomings and complications of the circulating ethical frameworks. He analyzed 22 major guidelines to find out “to what extent ethical objectives are actually implemented and embedded in the development and application of AI, or whether merely good intentions are deployed" (Hagendorff, 2020, p. 2). Even though both researchers used different approaches, they came to surprisingly similar conclusions as to why the frameworks are not as promising as they seem. The following four paragraphs will present their common ideas and elaborate on them.

### *Lack of accountability*

One of the shortcomings concerning the implementation of the ethical guidelines is a lack of accountability. The ethical frameworks are only suggestive and there is no punishment for those who do not work according to the guidelines. An example that

illustrates this is the frequent mention of the safeguarding of human autonomy in ethical guidelines (Fjeld et al., 2020; Jobin et al., 2019). For example, Google's own guidelines condemn "(T)echnologies whose purpose contravenes widely accepted principles of international law and human rights" (Google, 2018), of which human autonomy is one. This contradicts directly with the ways in which users' decisions can be, and are influenced online through psychological targeting using AI technologies (Matz et al, 2017). Organizations continue using these techniques because it is possible, lucrative and above all, legal. Except for ethical concerns, there are no real constraints in using these techniques and some form of accountability is necessary to provide some form of constraints.

## *Economic incentive*

The example above relates to another reason proposed by both authors, Hagendorff (2020) and Mittelstadt (2019), on why the ethical frameworks are not likely to meet their goal. Namely, the economic incentives that drive many AI projects. The Open Letter (Russell, et al., 2015) already mentioned that AI is used more frequently in economically valuable technologies boosting its funding and research. This causes the discussions on ethics of AI to have similarities with conventional business ethics where "ethical codes designate and defend social status and expertise more than enforce consistent moral or societal virtues" (Greene et al., 2019, p. 2124). In business, ethical considerations can be seen as a roadblock, barring the fastest way to make money. In this view, the practice of ethics is comparable to a game, where the objective is to get as close as possible to the moral boundaries. The moral boundaries are not yet established for AI though, therefore making the game very easy. Making the game harder by could be a solution, but the real problem is not the difficulty of the game, it is participating. from an ethical point of view, it is not desirable for the objective to be testing the moral limits to increase the amount of profit. For soft-law (*soft* because it is not *yet* legally binding) to become effective, the economic incentive should make place for something else, where ethics is helping to reach the goal rather than blocking the way.

## *Need for a common goal*

The first two issues, the lack of accountability, combined with the economic incentives, thus lead to a third concern surrounding the effectiveness of ethical guidelines, the absence for a *common goal* in the field of AI. Mittelstadt (2019), as stated before, makes a comparison with practitioners in health, where principled ethics is applied widely, and "soft-law" has proven to effect ethical behaviour in a positive way (Campbell & Glass, 2001). He argues that it is not just to generalize the working of this approach to AI, one of the reasons Mittelstadt gives for this, is that in health everyone shares the common goal of helping their patients, but in the AI sector there is no such established goal yet (Mittelstadt, 2019). Even though the ethical frameworks give a sense of a common goal, making AI beneficial for society, this is not what AI originally emerged from. The Darthmouth Conference was more about finding new *possibilities* within computer science (McCarthy et al., 2006), than finding new ways in which computer science could explicitly benefit society by creating AI. Apart from discovering new technological posibilities, AI technologies later also became a lucrative buisiness. For health practitioners, the common goal is clear, but within the AI community, this goal is still

divided. The lack of a clear common goal shared by *everyone* in the field of AI, resonates in the educational system. The focus lies more on the theory behind AI and its workings rather than the consequences (Eaton, et al., 2018). To elaborate on the parallel with health practitioners, it is as if medical students were taught the chemical workings of medicine in the body without getting told about the effects on the human taking them.

### *Lack of technical guidance*

Lastly, a more straightforward, but very important point of criticism when looking at the use of guidelines by developers, is the lack of practical guidance given by the frameworks surrounding ethical AI. A critical analysis of seven documents found that, “Other forms of expertise appear in these statements, but the problems themselves are to be solved by experts in the technical features of AI/ML systems" (Greene et al., 2019, p. 2129). It is thus stated that the technical experts are called upon to solve the problems that arise due to unethical AI. Going well with this finding is one of the eight common principles, Professional Responsibility, described in terms of accuracy, responsible design and consideration of long-term effects. This principle was found in 78% of the documents analysed by Fjeld et al. (2020). Even though these are recent papers, the Open Letter had already addressed developers for the need for research in how to develop AI systems that are robust and behave as intended in the third short-term research priority (Russel et al., 2015). In spite of these calls for action from developers of AI, both Mittelstadt (2019) and Hagendorff (2020) concluded that there is a lack of technical elaboration on how to implement the principles presented in the guidelines, while technical details could help developers to feel involved. Hagendorff (2020) is even suggesting that “the generality and superficiality of ethical guidelines in many cases not only prevents actors from bringing their own practice into line with them, but rather encourages the devolution of ethical responsibility to others" (Hagendorff, 2020, p. 14).

Corresponding with the concerns raised about the actual effects of the ethical frameworks is the result of a study done on the Association for Computing Machinery’s (ACM) code of ethics (McNamara et al., 2018) where it was found that “no statistically significant difference in the responses for any vignette were found across individuals who did and did not see the code of ethics, either for students or for professionals" (p. 732). They did a study with 63 software engineering students and 105 professional software engineers. After dividing the group in two and only giving one half the ACM code of ethics, they asked how they would behave in eleven situations with ethical complications. These situations were related to the following issues, *the responsibility to report, user data collection, intellectual property, code quality, honesty to customer to time* and *personnel management*. After concluding that it did not influence the choices of the participants, they called for future research on “identifying interventions that do influence decision making” (p. 732). This research assumes that programmers have an influence and responsibility on the future of AI, as it were their choices that were deemed useful to investigate. However, it did not go into detail as to *why* the code of ethics did not have an influence.

## ***Different approaches***

If these concerns turn out to be catastrophic for the ethical guidelines, it is wise to keep an eye open towards other attempts to secure a beneficial future for AI. Thilo Hagendorff

(2020) argues that “the prevalent approach of deontological AI ethics should be augmented with an approach oriented towards virtue ethics aiming at values and character dispositions". To understand this quotation, one must know the difference between the two types of ethical theories mentioned. Deontological ethics is, oversimplified, concerned with what we ought to do according to universal rules. It is doing ethics by following rules or laws. These rules and laws are to be seen apart from individual cases. Building an autonomous weapon would be considered bad by a deontologist who considers it a universal rule that *you shall not kill.* Virtue ethics is not concerned with universal rules or law but with the individual. It emphasizes a person’s good qualities or “virtues” (Goldsmith & Burton, 2017). So, the opportunity of building an autonomous weapon would raise the question “Do I want to be the kind of person to build such a weapon?”. A virtuous person would say “No” to this question. Bearing in mind that the documents surrounding the ethical debate on the creation of AI are termed *soft law*, it is not unreasonable to place the ethical guidelines within the deontological approach. Hagendorff (2020) says that a more virtuous approach in AI ethics will take away the “negative notion of ethics” and will “broaden the scope of action, uncovering blind spots, promoting autonomy and freedom, and fostering self-responsibility" (p. 114). Also saying that this should be accompanied by institutional changes from ethics in education to establishing institutions for complaints.

A more technical approach in ethics of AI is suggested by Iyad Rahwan (2017). He proposes an *algorithmic social contract* where society is not only *affected by,* but also *a part of* the AI systems. By giving various societal stakeholders the tasks “to identify the fundamental rights that the AI must respect, the ethical values that should guide the AI’s operation, the cost and benefit trade-offs the AI can make between various stakeholder groups, etc.” (p. 9), society is added to the system’s loop,. He calls this *society-in-the-loop* (SITL)*,* best illustrated with an example provided by Rahwan himself. Autonomous cars have almost become the present personification of the trolley problem (Thomson, 1976). Should algorithms of autonomous cars protect passengers over pedestrians or the other way around? SITL proposes to find answers by asking society as a whole and implement the outcomes into the systems. This would for instance reduce the possibility of such a technology causing an increase of inequality. People who can afford autonomous cars would for not be favored as passengers in such situations for example, unless it’s the conclusion of the *social contract* that has been agreed upon by all stakeholders who are affected by the technology. This way ethical guidelines are not controlling the ethical direction AI should take, but society. If society would decide that security is always more important than privacy, it would in SITL systems, in contrast to the ethical guidelines, be possible to skip the issue of privacy as a whole.

# 4. Method

The conducted research is constructed following the guidance provided by Ritchie and Lewis (2003) on qualitative research including but not limited to designing the research and setting up and carrying out a focus group. In their book they elaborated on common methods used and different approaches one can take in qualitative research. This research is set up as an applied policy research, which can be explained as: “The review process not only assesses the success or failure of the policy or procedure, it also encapsulates the implementation of these policies (…) the research is required to gather

specific information and has the potential to create actionable outcomes" (Srivastava & Thomson, 2009, p. 73). To put this in the perspective of this paper, the goal is to gather information on programmers' views on the implementation of ethical guidelines and maybe provide actionable outcomes if considered necessary.

Data was gathered through a focus group with eight participants who joined the group voluntarily and were not given any compensation. Three AI bachelor and two AI master students, one Robotics and Mechatronics master student, one Computer Science master student and an employee of CGI with experience in programming for AI projects were brought together in a Microsoft Teams meeting. All participants were Dutch and the group was homogenous in the sense that only people with a background in AI were invited. This is because ethicist and philosophers have proven to be busy with the moral questions of AI, but arguably the developers are called upon to implement the principles (Greene et al., 2019; Hagendorff, 2020; Mittelstadt, 2019). The participants had all replied to a message asking if they wanted to participate in a conversation about the development of AI, *not* mentioning the focus on ethical guidelines to prevent any bias. This message was posted in Dutch on LinkedIn, Facebook and sent through WhatsApp by the researcher in different groups surrounding AI and on the researchers own accounts. The spoken language of the meeting was Dutch, this was to eliminate the chance of people having difficulty to express themselves due to a lack of English-speaking capabilities. The exact message is available in appendix B. Apart from the participants, an independent minute taker and the host were present in the meeting.

The reason for a focus group rather than a survey is the objective to get a more thorough understanding of what developers of AI think about the guidelines and how they implement them in their work or study. The idea of individual interviews was also considered but a focus group got the upper hand "(B)ecause group discussions allow participants to hear from others, they provide an opportunity for reflection and refinement which can deepen respondents' insights into their own circumstances, attitudes or behaviour" (Ritchie & Lewis, 2003, p. 37). So, there is room for ideas that might not be expressed in a survey or individual interviews. The focus group was semi-structured and lasted one hour. The topic guide included the following topics: *implementation of ethical guidelines, the influence of the ethical guidelines* and *ways to increase the use of ethical guidelines*. The whole conversation was recorded with permission of all participants. The recording was transcribed intelligent, the transcribed interview is available in appendix A.

The data was analysed using a method based on the framework analysis steps for applied policy research provided by Ritchie and Spencer (1994). The analysis started with familiarizing with the transcript by reading it thoroughly. After getting familiarized, emerging themes or concepts that came forward in the data were gathered. From these concepts and themes, expressed by the participants and interpreted by the readers, categories were created to form a thematic framework. This was later used to classify the data. Because the researcher was familiar with the a priori issues, this was done by the researcher as well as an independent second reader. This way, the influence of the original research aims on the thematic framework was reduced, and issues raised by the respondents got proper attention. After both independently gathering themes and concepts, they were discussed and evaluated on the differences until a consensus on their meaning and relevance was reached. Then they were turned into categories. Parts -one or multiple sentences- of the data were identified and placed within the categories.

This was done both by the researcher and a third reader who was informed about the categories and what they stood for. The last step was dividing the categories into subcategories to arrange the extracted data into common ideas and thoughts of the participants within each category.

# 5. Results

The categories that were used to sort the themes were: *target audience, reasons for creating ethical guidelines, using ethical guidelines voluntarily, enforcing the use of ethical guidelines* and *encouraging the use of ethical guidelines.* The results are summarized in the rest of this chapter. An important thing to keep in mind while reading the results is that four out of eight participants knew about the ethical guidelines beforehand and one of these four could tell about the general contents of the guidelines.

## *Target audience*

The target audience of the guidelines was referred to in different levels starting with people being involved in producing AI and finishing with the whole world. The first group in this 'order' were the developers of AI and managers instructing them. Followed by companies, ranging from start-ups (using AI) to world leading tech-companies like Facebook and Google. It did not stop here as it was mentioned that "anyone with half a laptop can create a model in Python" (see appendix A, 46),[1] elaborating on a shared responsibility and thus increasing the target audience towards everyone with a laptop and some interest in programming. The largest group mentioned as a target audience was essentially everyone, including laymen. The provided argumentation was that the ethical guidelines can be used by everyone to see what is good and what is not. Guidelines can show everyone that it *is* possible to use and create AI in an ethically responsible way.

## *Reasons for creating ethical guidelines*

Understanding for the need of ethical guidelines was displayed by the participants accompanied with positive attitudes towards the idea of ethical principles for AI. The prevention of malicious use of AI was mentioned as an incentive for the ethical guidelines. However, the most protruding reason for creating ethical guidelines mentioned by the participants was minimizing unforeseen consequences in the use and production of AI. Also, the issue of money was brought up. The above-mentioned target group *companies* as well as people wanting to make money using AI were accused of ignoring ethical considerations. Stopping and preventing this to happen was given as another reason as to why the guidelines are needed. A more practical reason given by the group was the need for AI systems to be less of a 'black box' making important decisions that affect humans. A general agreement is needed according to some of the participants that AI systems must be made 'explainable' and guidelines might help to achieve this. The general view was that indeed *something* need to happen to raise ethical awareness and that ethical guidelines are a good idea to do this. However, the implementation and execution raised some concerns among the participants. Often reflecting in statements like, "Creating [ethical guidelines] is pretty realistic, sticking to it will be harder" (see appendix A, 50).

[1] All quotations from the focus group are translated from Dutch to English by the researcher.

### *Using guidelines voluntarily*

Before elaborating on the concerns, there were some ideas on why ethical guidelines might always be good to make regardless of them being compulsory. Some voluntary uses of guidelines were mentioned. It was for example, proposed that they could serve as a reference book for anyone who wants to see if the work they are doing is ethically responsible. Starting companies who want to stand out, could use it as a statement, not only promoting their product but also showing that they actively follow certain moral standards. It might raise goodwill for the product by its intended audience. Another idea that was put forward, was that journalists might use it at the end of their articles concerning problems of AI to show the people that there *are* ethical guidelines, but that these were disregarded.

### *Enforcing the use of guidelines*

Although voluntary use of ethical guidelines was considered an option, only *creating* the guidelines would not be enough according to most. A comparison was made between *global warming* and the development of AI: “I still think that guidelines on a voluntary basis won’t work. Everyone knows you should be good for the environment and look how that’s going" (see appendix A, 65). There was considerable support for the enforcement of using ethical guidelines to avoid a similar situation. One example was the use of an ethical “quality mark”. This implies that before using AI on a great scale or putting a product where AI is involved on the market, it should receive a quality mark that vouches for this product following the ethical guidelines of AI. Who should give this, or what exact guidelines to follow was not elaborated upon. Another example was more focused at the research community of AI, namely a mandatory pros and cons paragraph in all papers concerned with AI. This would force researchers to think about the consequences of their work and make sure that other people who read the article are informed by this as well. The most protruding way of enforcing ethical guidelines was by law and order. Introducing fines on a national as well as international (UN) level for anyone who does not work according to the ethical guidelines. It was discussed however that doing this would be very complicated and there was a doubt whether this would be possible. An argument against fines was that big companies using AI have enough money to pay fines, so the fines would have to be enormous to have any effect. Enforcing the guidelines through law and order was thus considered a very difficult task.

### *Encouraging the use of ethical guidelines*

Even though enforcing the guidelines was thought to be a difficult task, a step in the right direction might be taken in encouraging people to keep ethical considerations in mind while working on or with AI. The focus group provided many proposals that might stimulate this. Subsidies for research into the ethical side of AI was one of these, another suggestion was to raise awareness by having large informative campaigns aimed at a large public or giving warnings at the distribution of certain tools or information. “So every time you want to import and download Keras [deep learning software], that something, probably being skipped, a warning or something, will pop up to throw the guidelines in their faces and that you, in this way, maybe from the start already give a little push or provide help to learn something about it" (see, appendix A, 67). As well as the distribution of tools, the distribution of information was mentioned, putting these warnings at the start

of YouTube tutorials or Wikipedia pages about AI might initiate more ethical thinking. Together with doing this, it was agreed upon that AI ethics and the emerging guidelines should be a mandatory subject in all education that has anything to do with AI. One of the bachelor AI students, who did not have any mandatory ethics in his programme said that if ethical guidelines were to be taught thoroughly, “I would not take a job where I would have to throw everything I had learned out of the window and do something that I know is bad" (see appendix A, 73). Another student agreed with this and said that the Master AI involved some ethics, but she was astonished that fellow ex-students, who had already started working right after their bachelor, had never been taught any of it, while these people are the ones who “might do something silly" (see appendix A, 74). Finally there was mentioning of a “cultural shift”, stating that before something will change, something big must happen within the whole AI community. Someone said that if the big companies will start, the rest will follow. While another said that even a cultural shift will not be enough because there will always be immoral parties to stir things up.

# 6. Discussion

The research was mainly aimed at how the ethical guidelines of AI influence developers of AI. However, when looking at the results it is important to note that the conversation often went beyond the participants’ own experiences. For example, when speaking about the use of ethical guidelines there was often talk about how the guidelines *might* be used rather than participants’ own experiences. This is probably because most of the participants were students, they have not been part of big projects where AI is involved yet. Even though the results do not provide personal experiences in the use of ethical guidelines, developers of AI were mentioned as an important target group of the guidelines by the participants and the results therefore apply to developers of AI as well. Also, the participants agreed with the literature on the reasons why ethical guidelines are needed (Hagendorff, 2020; Fjeld et al., 2020). The background of the participants accounts for the main limitation of the research. The results could have been different if the sample had existed mostly of developers of AI from the work field. There were 8 responses to the posted message, who were all invited. Therefore, it was not an option to pick a more varying group of participants. To increase the chance of having this option, different ways of recruiting participants might be useful, contacting companies for example. A longer time period in which people can respond could also prove effective to find a more various group of participants.

The duration of the focus group meeting was one hour. This was chosen to increase the chance of volunteers enlisting for the meeting. One hour turned out to be a good estimation to discuss the planned topic guide. If more time had been available for the research however, more focus groups could have been organized with the same or different groups. An interesting topic to add to the topic guide for future or maybe longer focus groups could be the Society-In-The-Loop idea by Rahwan (2017) for example, which was discussed in the theoretical framework. It would be interesting to see what developers of AI think about his, more technical, approach. Furthermore, because this is a qualitative research, interpretation of the data plays a big role. Measures have been taken to reduce the interpretation bias during analysing the data, this resulted in fruitful debates on the nature of the categories and what to place within which category, successfully reducing, but not eliminating, biases.

A difference between the results and the literature is the lack of comments about the absence of technical detail in the ethical guidelines. This was probably because only one of them knew about the contents of the guidelines and thus had the knowledge to criticize them. However, the results do show that the guidelines might be aimed at various target audiences. This might account for the lack of technical detail in the documents, as laymen for example, have no need for technical details. A consequence of a lack of technical detail might be that developers feel excluded and lose interest. The intended target audience is thus an important topic for future research. Another interesting point visible in the results is that all participants agreed on the need for ethical steering within the development of AI, but the actual knowledge of research in AI ethics turned out to be minimalistic. Even though this is a qualitative research and it would not be grounded to generalize the results. It is at least worrying that they did not know more about the ethical approaches in AI, as they will likely be the ones working on AI projects after graduation. Future research ought to be conducted on finding out how developers of AI come in contact with ethics and how this could be increased.

The summarizing result, that came forward from the other results, is the need for a cultural shift in the AI community. For example, a ‘common goal’ with a non-economic incentive like that of health practitioners, as suggested by Mittelstadt (2019), shared by every AI developer. AI is often used to make money with apparently the end justifying the means. If there is money to be made, suggestive ethical frameworks will not be enough. Therefore, enforcement is needed in the form of laws and research regulations. However, this is a hard objective and might therefore not happen anytime soon. While waiting for legislation, like Hagendorff (2020) said, a more virtuous approach to AI ethics is needed. This might be a good way to start From the focus group it was clear that there are ways to raise ethical awareness that are not utilized today. Instead of waiting for the lawmakers to achieve the immense difficult task of making laws for AI, this paper proposes, based on the literature and the results, to put more focus on personal ethical awareness surrounding AI. Starting with the AI community, but not stopping there as AI is going to impact the whole world. It is a small percentage of the population that is creating a technology that will influence all the rest. To help the ethical guidelines find their way into the common image of all that has to do with AI, ethical awareness must be intertwined with the technology. This is important to realize for both parties, developers and those who are affected by the technology. The focus group and literature provided insight in how to start raising ethical awareness surrounding AI technologies, paving the way for (mandatory) ethical guidelines and legislation, giving rise to the following proposals:

- **Making ethics mandatory in all AI and AI related studies.** Arguments supporting this have been made as well as proposals on how to approach ethics in the AI curriculum as well (Eaton, et al., 2018; Furey & Martin, 2019; Goldsmith & Burton, 2017). By doing this, the step to asking ethical questions about projects in meetings, to colleagues or oneself will become easier. Ethical guidelines for AI could be used to give shape to these courses.
- **Giving warnings and food for thought at the distribution of AI-tools and information about AI.** People who did not have an education in AI will be helped to learn about the ethical side of AI and the impact their work might have. Ethical

guidelines for AI could be displayed like a Terms of Use agreement, that have to be read and accepted before continuing.

- **Organize informative campaigns to let *everyone* know about the good AI has to offer but also the dangers it brings.** By doing this companies might feel more pressured to actively follow the rules they claim to follow. This will also make the general public more involved in securing a "beneficial" future for AI.
- **Taking responsibility by everyone wherever they can in helping to raise ethical awareness surrounding AI.** Developers of AI must realize that they potentially influence millions of people with their programs and show that they take the responsibility that comes with this influence. Incidents of employees going against their employers in the AI industry due to ethical concerns could serve as an example (Cameron & Conger, 2018; Tiku, 2019). Not only developers of AI but also people who are affected by AI should be encouraged to ask critical questions and talk about the subject whenever they feel it could help in raising ethical awareness, for example by beginning a conversation about the ethical guidelines for AI.

# 7. Conclusion

The aim of this paper was to find out what the influence is of ethical frameworks, created to guide the ethical development of AI, on developers of AI, according to developers of AI. This research has provided limited results to adequately answer this question. However, it has been able to identify areas that are interesting for future research that relate to the sub questions. Concerns about the effectiveness of ethical frameworks have been identified both within the literature and the data. These are: the economic incentive within the AI development overruling ethics, the need for enforcement of the guidelines or other ways to hold developers accountable and the need for a common goal or a cultural shift that could be stimulated through information campaigns and education. Future research could focus on the legitimacy of these limitations and how to tackle them. Tackling these limitations will provide ways to improve the influence of ethical guidelines on developers of AI. In the discussion section, four actions are proposed to start raising more ethical awareness in the world of AI to pave the way for (mandatory) ethical guidelines and legislation.

# 8. References


Cameron, D., & Conger, K. (2018, June 3). *Google Is Helping the Pentagon Build AI for Drones*. Retrieved from Gizmodo: https://gizmodo.com/google-is-helping-the-pentagon-build-ai-for-drones-1823464533

Campbell, A., & Glass, K. (2001). The Legal Status of Clinical and Ethics Policies, Codes, and Guidelines in Medical Practice and Research. *McGill law journal, 46*, 473-89.

Clement, J. (2020, July). *Number of social media users worldwide 2010-2021*. Retrieved October 30, 2020, from Statista: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users

Eaton, E., Koenig, S., Schulz, C., Maurelli, F., Lee, J., Eckroth, J., . . . Williams, T. (2018). Blue sky ideas in artificial intelligence education from the EAAI 2017 new and future AI educator program. *AI Matters , 3*(4), 23-31. doi:https://doi.org/10.1145/3175502.3175509

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020, January 15). Principles Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *Berman Klein Center Research Publication No. 2020-1*. doi:http://dx.doi.org/10.2139/ssrn.3518482

Furey, H., & Martin, F. (2019). Introducing Ethical Thinking About Autonomous Vehicles Into an AI Course. *AAAI*, 7900-7905. Retrieved from https://pdfs.semanticscholar.org/995c/5c85dc26bd1f2d77e273d58705fc0777ad9c.pdf

Goldsmith, J., & Burton, E. (2017). Why teaching ethics to AI practitioners is important. *ACM SIGCAS Computers and Society*, 110-114.

Google. (2018). *Artificial intelligence at Google: Our principles*. Retrieved October 10, 2020, from https://ai.google/principles/

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, (pp. 2122-2131). doi:10.24251/HICSS.2019.258

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines*, 99-120. doi:https://doi.org/10.1007/s11023-020-09517-8

Harari, Y. N. (2017). Reboot for the AI revolution. *Nature, 550*(7676), 324-327. Retrieved October 20, 2020, from https://nature.com/news/reboot-for-the-ai-revolution-1.22826

High-Level Expert Group on AI. (2019). *A Definition of AI: Main Capabilities and Diciplines.* European Commission.

High-Level Expert Group on AI. (2019). *Ethics Guidelines for Trustworthy AI.* European Commission.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat Mach Intell*(1), 389-399. doi:https://doi.org/10.1038/s42256-019-0088-2

Johnson, K. (2019, July 19). *How AI companies can avoid ethics washing*. Retrieved October 17, 2020, from VentureBeat: https://venturebeat.com/2019/07/17/how-ai-companies-can-avoid-ethics-washing/

Joseph, G., & Lipp, K. (2018, September 6). *IBM Used NYPD Surveillance Footage to Develop Technology that Let's Police Search by Skin Color*. Retrieved October 20, 2020, from The Intecept: https://theintercept.com/2018/09/06/nypd-surveillance-camera-skin-tone-search/

Lapowsky, I. (2019, March 17). *How Cambridge Analytica Sparked the Great Privacy Awakening*. Retrieved October 20, 2020, from Wired: https://www.wired.com/story/cambridge-analytica-facebook-privacy-awakening/

MarketsandMarkets. (2018). *AI in Social Media Market by Technology, Application, Component, Enterprise Size, End-User, and Region - Global Forecast to 2023.* Retrieved from https://www.marketsandmarkets.com/Market-Reports/ai-in-social-media-market-92119289.html

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017, November 28). Psychological targeting as an effective approach to digital mass persuation. *PINAS*, 12714-12719. doi:https://doi.org/10.1073/pnas.1710966114

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine, 27*(4), 12-14. doi:https://doi.org/10.1609/aimag.v27i4.1904

McNamara, A., Smith, J., & Emerson, M.-H. (2018). Does ACM's code of ethics change ethical decision making in software development? *ESEC/FSE 2018: Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (pp. 729-733). New York: Association for Computing Machinery. doi:https://doi.org/10.1145/3236024.3264833

Mittelstadt, B. (2019). Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence, 1*(11), 1-7. Retrieved October 5, 2020, from https://nature.com/articles/s42256-019-0114-4

Rahwan, I. (2017). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, 5-14. doi:http://dx.doi.org/10.1007/s10676-017-9430-8

Ritchie, J., & Lewis, J. (2003). *Qualitative Research Practice: A Guide for Social Science Students and Researchers.* London: SAGE Publications Ltd.

Ritchie, J., & Spencer, L. (1994). Qualitative data analysis for applied policy research. In A. Bryman, & R. G. Burgess, *Analyzing qualitative data* (pp. 173-194).

Russel, S., Dewey, D., & Tegmark, M. (2015, 12 31). Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine, 36(4)*, 105-114. doi:https://doi.org/10.1609/aimag.v36i4.2577

Russell, S., Dietterich, T., Horvitz, E., Selman, B., R. F., Hassabis, D., . . . Phoenix, S. (2015, 12 31). Letter to the Editor: Research Priorities for Robust and Beneficial Artificial Intelligence: an Open Letter. *AI Magazine, 36(4)*, 3-4. doi:https://doi.org/10.1609/aimag.v36i4.2621

Srivastava, A., & Thomson, S. B. (2009). Framework Analysis: A Qualitative Methodology for Applied Policy Research. *JOAAG, 4*(2), 72-79. Retrieved from https://www.researchgate.net/publication/267678963_Framework_Analysis_A_Qualitative_Methodology_for_Applied_Policy_Research

Thomson, J. J. (1976, April 1). Killing, Letting Die, and The Trolley Problem. *The Monist, 59*(2), 204-217. doi:https://doi.org/10.5840/monist197659224

Tiku, N. (2019, July 7). *Most of the Google Walkout Organizers Have Left the Company*. Retrieved October 30, 2020, from Wired: https://www.wired.com/story/most-google-walkout-organizers-left-company/

Zeng, Y., Lu, E., & Huangfu, C. (2018, December). Linking Artificial Intelligence Principles. *arXiv e-prints*. Retrieved October 30, 2020, from https://ui.adsabs.harvard.edu/abs/2018arXiv181204814Z

# 9. Appendices

## *Appendix A*

Focus group transcript.

*Iedereen verwelkomen en bedanken voor hun komst. De volgende regels worden besproken:*
*Microfoon op mute, tenzij je iets wil zeggen. Hand omhoog wanneer je wil spreken en als de host je de beurt geeft mag je spreken. Iedereen mag zeggen wat hij wil en wordt serieus genomen. Wanneer iedereen toestemming heeft gegeven wordt de opname gestart.*

1. Host: Om te beginnen wil ik vragen of iedereen zichzelf kort kan introduceren. Vermeldt daarbij je naam, wat je doet en je motivatie om betrokken te zijn bij KI.

2. Bachelor Student AI, Utrecht, 2: 3e jaar bachelor KI UU. Ik vind het leuk dat het een vooruitstrevend gebied is, je kan er veel kanten mee op. Alle bedrijven hebben het nodig, we zijn pas bij het begin. Dus daarom heb ik er wel interesse in.

3. Employee CGI, Programmer: Ik ben al 2 jaar aan het werk als HI engineer en implementeur. Ik vind AI bijzonder interessant puur om het volgende, stel je hebt in de wiskunde een bepaald algoritme of een bepaalde formule waarmee je iets kan platsslaan of iets kan uiten, sommige systemen kennen we niet goed genoeg om dat te kunnen doen waarbij AI eigenlijk gewoon zoveel mogelijk variabele neemt om dat wél te kunnen doen zonder dat we daar per se de formule voor kennen. Dat is voor mij best wel sterk, je hoeft het systeem niet per se te begrijpen en je kan het overlaten aan de data zelf.

4. Master Student Robotics, Twente: Ik doe een Master robotica en mechatronica aan de UT. Ik ben vooral geïnteresseerd in reinforcement learning en zelflerende algoritmen voor het leren van beweging en het oplossen van taken. Vooral die kant, robots die leren lopen en navigeren etc. Ik vind het een magisch iets dat het begint te kunnen de laatste paar jaar. Echt heel recent dat het in robots iets kan doen, dat vind ik leuk. Ik heb nu 2 jaar een burn-out waardoor ik niet aan het studeren ben, maar ik ben een beetje aan het opkrabbelen. Ik ben er nu voorzichtig mee bezig een paar uur per dag en leuk dat ik erbij mag zijn.

5. Bachelor Student AI, Utrecht, 1: Ik ben een 3e-jaars KI student aan de UU. Wat mij aantrok in de studie KI was dat het een combinatie is van informatica, psychologie en filosofie. Dat is ook de reden dat ik het hier in Utrecht ben gaan doen

6. Bachelor Student AI, Nijmegen: Ik studeer AI in Nijmegen. Wat me daar aan aantrekt is dat het een mooie combinatie is van psychologie en techniek zonder dat het te technisch wordt. Dat vind ik er wel mooi aan.

7. Master Student AI, Utrecht, 1: Ik ben eerstejaars master student KI, ik heb eerst de bachelor gedaan ook in Utrecht. Ik sluit me eigenlijk een beetje aan bij Bachelor Student AI, Utrecht, 1 en Bachelor Student AI, Nijmegen. Wat mij heeft getrokken bij Ki is dat het best wel breed is, zowel techniek als filosofie en psychologie. Ik vond tijdens de bachelor ook vooral de logica kant heel erg leuk. In Utrecht had je dan de kans om heel breed te studeren in KI. Dat is gewoon cool dat dat allemaal kan.

8. Master Student AI, Utrecht, 2: Ik ben een eerstejaars masterstudent AI in Utrecht. Ik ben destijds begonnen aan de bachelor KI omdat die mengeling van cognitieve neurowetenschap, psychologie en informatica mij heel erg trok. En gaandeweg ben ik erachter gekomen dat alle andere pijlers van KI mij ook heel erg trokken en daarom ben ik door gegaan met de master AI.

9. Master Student Computer Science, Utrecht,: Ik heb ook de AI-bachelor gedaan en ben toen meer richting de informatica gegaan met een master informatica in Utrecht. Nu doe ik een specialisatie in de richting van wetenschapscommunicatie in Leiden. Wat mij aan KI trok, een beetje aansuitend op de rest, is de veelzijdigheid. Nu ben ik wat meer aan het kijken hoe je KI of een andere magische wetenschap aan het brede publiek kan communiceren.

10. Host: Hebben jullie voorbeelden, uit eigen ervaring of die van een ander, waarbij KI-code een ander effect had dan verwacht? Twee minuten denktijd, eerst eigen ervaring en dan de ander.

11. Employee CGI, Programmer: Wat bedoel je met KI-code?

12. Host: Dat is aan jullie, je mag de vraag zelf interpreteren.

13. Na twee minuten.

14. Employee CGI, Programmer: Als ik een code schrijf dan werkt hij naar behoren. Ook als er een error is. dus daar blijft het bij

15. Host: Dus wel een fout in de code?

16. Employee CGI, Programmer: Ja.

17. Master Student AI, Utrecht, 1: Ik heb eigenlijk twee dingen bedacht. In een eerste jaars vak moest je iets met NLP testen, moest in Jupiter notebook features bedenken. Zo hoog mogelijke score met F score en accuracy ofzo. Ik had bedacht dat je een score van 100 kon krijgen als je zo veel mogelijk features

zou toevoegen. Had niet verwacht dat features tegen elkaar in zouden gaan etc. Dat had ik niet verwacht van bepaalde KI code. Tweede ding: net logo met ias, mier laten rondlopen. Begonnen als simpele opdrachten, prisoners dilemma. Daaraan had ik niet verwacht dat je eerst een vrij simpele opdracht kreeg en dat er iets emergent uitkwam.

18. Master Student Robotics, Twente: Ik had een bot gemaakt en deze moest spelen en dan tegen oudere versies van zichzelf moest om steeds beter te worden. Maar dat ging niet zoals verwacht, hij werd niet telkens beter maar het verschilde heel erg of hij goed was. Was geen pijl op te trekken. Dat was onverwachts dat self-play nog best ingewikkeld was. Maar nog iets flitsender: toen ik bezig was een gesimuleerde robotarm een doel te laten bereiken (zat vast aan de onderkant, met obstakels zoals een pijp met een bocht erin). Hadden we in elkaar gezet, gingen draaien, maar hij pleegde z.s.m. zelfmoord. Een paar keer achter elkaar. Wij dachten huh? Algoritme getest, maar alles werkte zoals het zou moeten. Maar toen we er beter over na gingen denken waren de functies zo ingesteld dat hij altijd gestraft werd, hij wilde gewoon zodat het zo snel mogelijk over was.

19. Bachelor Student AI, Nijmegen: Ik heb geen eigen ervaring, behalve als code niet werkt. Maar paar jaar geleden een bot op twitter, die leerde door met mensen te tweeten en tweets te lezen. Hij was volgens mij van Microsoft. Binnen een paar uur was die bot heel racistisch geworden en begon hij scheldwoorden te gebruiken, hij is toen offline gehaald.

20. Host: Employee CGI, Programmer wat was niet duidelijk?

21. Employee CGI, Programmer: ik hoor code en niet model. Iedereen antwoord op de vraag model. Mijn code draait gewoon, dus werkt het gewoon en word ik niet verrast.

22. Host aan iedereen: geeft Bachelor Student AI, Nijmegen wel antwoord op de vraag?

23. Master Student Robotics, Twente: Ik zou het zelf anders interpreteren, model is alles wat je code is en je code zit in je model. Dus je ka je altijd verbazen door zelflerende trucjes

24. BEGIN:

25. Host: Dit was om in te komen, mijn bachelor gaat over ethische richtlijnen. Wie weet dat er ethische richtlijnen/principes voor het ontwikkelen van KI bestaan?

26. Handjes: Master Student AI, Utrecht, 1, Master Student AI, Utrecht, 2, Master Student Computer Science, Utrecht,, Master Student Robotics, Twente

27. Host: Wie van de mensen die ze kennen weten ook wat erin staat?

28. Alleen Master Student AI, Utrecht, 1.

29. Host: heb je ze gebruikt wanneer en waarvoor?

30. Master Student AI, Utrecht, 1: Ik volg momenteel een vak en dat is methods of AI research, daar wordt benadrukt dat alles wat je kan programmeren hoeft je niet per se te programmeren, let op wat je maakt. Ze vertelde dat er Microsoft guidelines zijn voor human AI interaction. Van dat soort guidelines moesten we toen een soort kleine opzet maken. Toen moesten we kijken naar een paar programma's en of deze zich wel aan de guidelines hielden. Ik heb ze dus niet zelf gebruikt bij het programmeren van een opdracht maar ik heb er wel mee gewerkt.

31. Host: Kan je uitleggen wat deze richtlijnen zijn?

32. Master Student AI, Utrecht, 1: Dat zijn een aantal guidelines ik ken ze niet precies maar een van de dingen is dat als er in een AI model iets fout gaat dat er dan gezien kan worden wat er is misgegaan. Dat er bijvoorbeeld genoeg feedback is, ook feedback verwerkt vanuit de mens. Dat is dan volgens mij ook weer ingedeeld in verschillende lagen van wat je met AI wil bereiken en hoe belangrijk zo'n systeem is.

33. Host: Ik zal een beetje aanvullen. Er zijn inmiddels al 84 richtlijnen met onderwerpen als, international human rights, promotion of human values, professional responsibility, human control of technology, fairness and non-discrimination, transparency and explainability, safety and security, accountability, privacy. Dus dit zijn richtlijnen die gebruikt kunnen worden bij het maken van KI. Waarom vinden mensen het nodig om zulke richtlijnen/documenten te maken?

34. Employee CGI, Programmer: Klein voorbeeldje: ik heb een tijdje voor een bank gewerkt en voor een verzekeringsbedrijf. Daar heb je vaak een model dat je risico gaat inschatten. Je wil daarbij weten waarom een risico ergens aangehangen wordt. Dit wil je heel duidelijk kunnen vertellen aan een persoon of bedrijf waarom dat risico eraan wordt gehangen. Dat is bijvoorbeeld een stukje explainable AI. Kijken waarom je een model een bepaalde keuze laat maken. En daarom is het van belang, want je kan niet zomaar zeggen van, dit hebben we gedaan omdat het model het zegt. Je moet het kunnen verklaren, anders ben je aan het discrimineren zonder dat je dat misschien weet.

35. Master Student Robotics, Twente: Ja super breed natuurlijk, misstanden voorkomen in elke zin. Het kan op allerlei manieren fout gaan, dus als je de kans daarop minder kan maken met een goed plan dan lijkt mij dat een goed idee.

36. Host: Zijn jullie het ook eens met het idee om principes op te stellen om zo die kans te verkleinen?

37. Master Student AI, Utrecht, 1: Ja ik vind het super goed dat dit soort richtlijnen worden gemaakt en ik zou dit bij elke AI bachelor willen terugzien. Het is zo makkelijk om iets te maken waar bias in zit, waar racisme in zit, waar discriminatie in zit als je het hebt over machine learning modellen. Het is zo makkelijk om per ongeluk iets te maken met een bepaald gedrag waarvan je niet meteen weet of dat uiteindelijk de bedoeling is. Dan heb je inderdaad iets van explainability nodig, of iets van die andere thema's. Dat daar iets van richtlijnen worden opgesteld van "joh dit is eigenlijk de bedoeling dat je je hier en hier aan houdt en dat dat vanaf het begin al kan worden meegegeven. Dat als er een nieuw project wordt gestart, dat er dan gekeken kan worden, houden we ons wel aan de standaarden van AI? Of zijn we maar wat aan het programmeren in de hoop dat we wat centen kunnen verdienen maar zijn we ondertussen echt veel te veel mensen hun privacy rechten aan het schenden. Dat was ook het probleem toen op een gegeven moment de AVG inging, dat was huilen binnen alles in de informatica. Big Data kon niet meer en het was het einde van een enorm groot iets binnen de KI. Maar tegelijkertijd is het ook eigenlijk wel top dat dat nu niet meer kan want je kan er zoveel foute dingen mee doen. Bedrijven kunnen je data scrapen en dat dan weer tegen je gebruiken. Dus in dat opzicht denk ik zeker dat het goed is dat er iets van ethische richtlijnen zijn waar bedrijven/projecten/studenten zich aan moeten houden.

38. Host: Employee CGI, Programmer, heb jij er wel eens belemmeringen door meegemaakt in je werk?

39. Employee CGI, Programmer: Nee niet bepaald. In het geval van privacy: je anonimiseert data. Je moet in je achterhoofd houden dat je iets moet kunnen uitleggen. Maakt het alleen maat interessanter om ermee te kunnen werken, je werkt met een diepere laag waar je van tevoren al over na had moeten denken.

40. Host: Op wie zijn deze richtlijnen gericht volgens jullie?

41. Bachelor Student AI, Utrecht, 1: Ik denk vooral op mensen die er heel snel en heel veel geld mee willen verdienen, aan KI. Die gewoon niet goed nadenken over wat ze doen en gewoon snel geld willen verdienen.

42. Master Student AI, Utrecht, 2: Ik wil het wat breder trekken want ik denk dat die gericht kunnen zijn op iedereen die iets te maken kan of wil hebben met AI. Natuurlijk onstaan er problemen door mensen die niet nadenken over wat ze doen en wat ze opzetten maar ik kan me ook heel erg goed voorstellen, zoals blijkt uit de eerste vraag eigenlijk al, dat er genoeg dingen redelijkerwijs niet echt voorzien hadden kunnen zijn. Dus ook voor dingen die er niet direct uitspringen

als zijnde een probleem je eigenlijk gedwongen wordt om alles te dubbelchecken of het niet op een aangegeven richtlijn manier problematisch is.

43. Master Student Robotics, Twente: Uiteraard voor de ontwerpers, maar ik hoop ook voor het management zodat die ook inzicht krijgen in wat wil je en wat wil je niet. Dat het niet alleen aan de ontwerpers is binnen zo'n bedrijf. Dat lijkt me wel goed.

44. Host: Dus het is een gedeelde verantwoordelijkheid?

45. Master Student Robotics, Twente: Ja zou ik wel zeggen

46. Master Student AI, Utrecht, 1: Ik denk ook dat het een gedeelde verantwoordelijkheid is. Grote bedrijven zoals google moeten zich ook eraan houden, maar iedereen met een halve laptop kan in python een model creëren. Voor dat soort mensen is echt net zo goed belangrijk om het te beseffen. Je zou misschien data kunnen anonimiseren door namen weg te halen, maar misschien is dat helemaal niet genoeg. Dus ook voor kleine startups.

47. Master Student Computer Science, Utrecht,: Om het nog iets breder te trekken, ik denk dat heel veel mensen er ook baat bij kunnen hebben. Niet alleen de mensen die het gebruiken en de makers, maar ook een leek op straat kan er baat bij hebben om te zien wat wel en niet goed is, en te zien dat het wel mogelijk is om dit op een goede manier te doen.

48. Bachelor Student AI, Utrecht, 2: Ik wil de link leggen tussen de richtlijnen van KI en de richtlijnen van onderzoeken. Goede onderzoeken moeten voldoen aan richtlijnen voordat ze gepubliceerd kunnen worden. Op dezelfde manier moeten we omgaan met KI-algoritmes, die moeten dan voldoen aan de richtlijnen voordat ze gebruikt worden bij bedrijven.

49. Host: die richtlijnen bij onderzoeken zijn best wel strikt. Hoe reëel is het om dit bij de KI te maken?

50. Master Student Computer Science, Utrecht,: Maken is best reëel, is al gedaan, maar het eraan houden wordt moeilijker. En bij het maken is misschien het moeilijk om het erover eens te worden. Over het gebruik ervan ik denk dat dat moeilijker zal zijn. Er zal een soort van cultuurverschuiving moeten komen om dat op grote schaal toe te passen.

51. Master Student AI, Utrecht, 2: Ik geloof dat Master Student Computer Science, Utrecht, zegt klopt, dat deze richtlijnen al bestaan, in elk geval omtrent machine learning. Bepaalde research-groepen moeten bijvoorbeeld bij het publiceren kunnen uitleggen aan commissiegroepen hoe dit ‘voor evil’ gebruikt zou kunnen worden. Zodat je er heel bewust mee om moet gaan en moet nagaan of het ook voor hele negatieve doeleinden kan worden gebruikt

52. Host: Wat maakt dan dat zo’n richtlijn goed is? Hoe is het goed te implementeren?

53. Master Student AI, Utrecht, 1: In dat opzicht is het een splitsing tussen wetenschappelijke richtlijnen en richtlijnen die in de praktijk moeten worden gebruikt. Je moet het best kunnen reguleren in de wetenschap dat je net als een conclusie en een discussie, een pros en cons paragraaf moet hebben en anders wordt het niet geaccepteerd in een wetenschappelijk blad. Ik denk dat op die manier daar onderzoeken best wel naar willen dansen dan is dat de manier hoe je ethisch bewust wordt van de programma's die je maakt. Ik zie niet gauw voor me hoe je dit in de praktijk bij bedrijven moet doen, dat je ze moet verplichten om dingen op hun website te zetten, "dit is hoe wij hierover na hebben gedacht met betrekking tot deze grote thema’s." Dan krijg je zoiets als met privacy verklaringen, dat zou kunnen maar ik zie niet voor me dat dat in een keer verplicht kan worden. In ieder geval goed om overal aan te bieden wat makkelijk te vinden is en kan dienen als naslagwerk.

54. Master Student Computer Science, Utrecht,: Ik ben het met Master Student AI, Utrecht, 1 eens dat verplichten in dit gebied niet de makkelijkste weg zal zijn. Meeste grote bedrijven staan boven de wet (of denken ze). Ik hoop dat er door een bewustzijn/cultuur verandering zal komen, ik zie toch wel vaker artikelen met "oh dit en dit bedrijf discrimineert" kortom, grote, grote ophef, als er dan iets veranderd, laat er maar ophef zijn. En als richtlijnen dan onder aan deze artikelen kunnen staan zouden ze misschien vanuit deze kant effect hebben.

55. Bachelor Student AI, Nijmegen: Ik denk dat het hanteren van zulke richtlijnen juist bij kleine bedrijven een probleem zal zijn. Die hebben zelf minder resources om zich actief aan zulke richtlijnen te houden, en zijn in tegenstelling tot grote bedrijven die als mooie prooi worden gezien, het vaak niet waard voor een waakhond o.i.d. om erachter aan te gaan.

56. Host: Wat bedoel je met een waakhond?

57. Bachelor Student AI, Nijmegen: Ik weet niet precies hoe die heten maar van die instanties gebonden aan overheden of de EU die bij inbreuken van dergelijke privacy-richtlijnen en dus misschien ook AI die bedrijven daadwerkelijk voor de rechter slepen en dergelijke boetes uitdelen.

58. Host: Zulke handhavers zijn er dus al wel volgens jou?

59. Bachelor Student AI, Nijmegen:Ja ik lees vaak genoeg in het nieuws dat bedrijven een boete krijgen omdat ze bijv. discriminerende keuzes hebben gemaakt.

60. Master Student AI, Utrecht, 1: Ik denk dat het handhaven door middel van boetes inderdaad heel lastig is, net als bij de wet AVG. Bijvoorbeeld dat werd gezegd over kleine bedrijven van "ja maar wie gaat er nou bij zo'n klein bedrijf kijken of er persoonsgegevens worden geschonden en dat is gewoon heel moeilijk te handhaven. Maar andersom, het kan juist iets zijn waarmee je je als startup zou kunnen verkopen door naast hun idee te zeggen dat ze zich ook nog aan alle richtlijnen houden Omdat ik wel denk dat het inderdaad moeilijk is om te handhaven denk ik dat het misschien eerst uit een collectief van grote bedrijven komen. Dat Google en Facebook zeggen hoe ze zich aan de richtlijnen houden want ik kan me voorstellen dat het dan een standaard wordt om dat ook te doen. In plaats van kleine bedrijfjes en studenten die dat doen dat het als cultuur de standaard wordt en dat kan misschien het beste als de grootste projectontwikkelaars dat eerst doen.

61. Bachelor Student AI, Nijmegen: Daar ben ik het mee eens, maar hoeveel mensen zich er ook vrijwillig aan zouden houden, er zijn altijd startups die snel geld willen verdienen en vrijwillige richtlijnen achter zich laten. Als dit geen negatieve consequenties voor henzelf heeft, dus dat ze er puur geld mee kunnen verdienen, dan is de enige oplossing die ik er dan voor zie is dat geen enkele klant meer producten zou willen waarvan niet bewezen is dat ze ethisch te verantwoord zijn. Maar dat verwacht dan een enorme shift in de maatschappij denk ik.

62. Master Student AI, Utrecht, 1; Ja maar dan krijg je dus inderdaad dat het misschien invoeren van dit soort regels als een manier voor projectontwikkelaars kan dienen om zich te laten kennen en te laten zien van "joh hier hebben wij ons wel aan gehouden", als een soort van bonus, dat is dan in ieder geval wel iets, dan lok je het gedrag uit om het als een extra iets te doen. Als je het helemaal niet hebt wordt het in ieder geval niet nageleefd. Je geeft ze dan een extra incentive om het wel te doen.

63. Master Student AI, Utrecht, 2: Ik heb eigenlijk twee reacties, als je het gaat gooien op een vrijwillige morele cultuurshift, immorele partijen gaan daar niet in mee. Als genoeg mensen hier geld uit willen halen gebeurt er alsnog heel veel naars. En een argument tegen opleggen van grote boetes, het kan misschien een medium-sized startup bankroet krijgen, maar zolang je het niet hebt over boetes die bijvoorbeeld relatief zijn aan de jaarlijkse inkomsten van bedrijven krijg je wel een situatie waar misstappen onder betaling nog steeds mogelijk zijn voor grote bedrijven. Dan is het voor de "rijke" nog een soort van legaal. Je moet ervoor betalen, maar het mag wel.

64. Host: Ze zijn dus met richtlijnen bezig, maar over de invloed wordt nog getwist. Zijn er ook andere manieren die wel zouden helpen?

65. Bachelor Student AI, Nijmegen: Ik denk nog steeds dat op vrijwillige basis richtlijnen hebben niet gaat werken. Iedereen weet dat je goed voor het milieu

moet zijn maar kijk naar hoe dat gaat. Vanuit grote instanties zoals de VN moeten richtlijnen worden opgesteld, deellanden moeten worden verplicht dit soort richtlijnen in hun wetten op te nemen. Binnen internationale instanties, als een overheid zelf AI onverantwoord in zou zetten, ook via internationale instanties sancties zou moeten krijgen.

66. Bachelor Student AI, Utrecht, 2: Ik had eigenlijk het zelfde antwoord. Denk dat er veel fouten gemaakt moeten worden voordat de goede dingen de kop op steken. Grote richtlijnen enige manier of grote teams van mensen die codes gaan analyseren, maar dat lijkt me niet realistisch.

67. Master Student AI, Utrecht, 1: Ik denk dat het met bedrijven die iets kwaad willen het altijd wel een kat en muisspel blijft tussen hoe hard je die richtlijnen moet doen en hoe je dat kunt handhaven. Maar misschien zou je het ook kunnen doen door het bij de kern aan te pakken, bij de distributie van bepaalde tools, dat je het daar wat moeilijker maakt. Dus elke keer als je Keras (deep learning software python) wil importeren en downloaden, dat er dan, iets dat waarschijnlijk ook wordt geskipt en niet doorgelezen, een manier komt, een warning of iets om nog een keer die richtlijnen in het gezicht te duwen. En dat je dan op die manier misschien vanaf het begin al een zetje of een beetje hulp geeft om ze er iets over te leren.

68. Master Student Robotics, Twente: Ik ben voor regulaties en wetten, vond klimaatverandering een goed voorbeeld. Vrijwillige dingen heb je altijd bad actors en dat hou je waar geld in het spel is. Ik denk dat je moet het zien zoals vliegtuigen of auto's waar ook bepaalde veiligheidseisen om minder mensen dood te laten gaan zoals gordels. Die moeten er ook zijn. Ik zou het heel dom en onverantwoord vinden als we als wereld zeggen we gaan allemaal belangrijke en dure dingen, daar regels voor maken maar we laten KI een beetje in een zwart gat, een soort wilde westen te laten. Terwijl het super krachtig gaat zijn en het gaat alleen nog maar krachtiger worden omdat het ook banen gaat vervangen enz.. Het lijkt me heel jammer als we daar geen regels voor gaan verzinnen alleen welke regels is wel een heel lastig punt, daar moet je wel overeenkomen. Ik hoor net een overkoepelend iets als de VN ofzo maar het is nog maar kijken hoe dat zou uitpakken. Ik had ook nog een vraag, wil je het nog hebben over de lange termijn van de KI, wanneer we het menselijk denkniveau overschrijden, want daar vindt het leuk om het over te hebben. Want wat ik daarvan vind is dat ik er volledig ervan overtuigd ben dat we daar wetten voor moeten hebben voordat de techniek er is omdat het een leven of dood situatie is. Meerde mensen hebben het al een existentieel risico genoemd. We kunnen allemaal doodgaan als iemand het verneukt.

69. Host: Heel interessant inderdaad maar net wat te groot om erbij te nemen dus daarom hou ik dat nog even terzijde.

70. Master Student Computer Science, Utrecht,: Ik zat zelf te kijken van wat er mogelijk is want het is ook niet allemaal heel makkelijk om dingen netjes, goed, niet privacy-technisch en niet discriminerend te maken. Dus misschien meer funding met onderzoek meer op dit soort initiatieven te gooien, startups belonen als ze er onderzoek naar doen. Dus iets meer vanuit de onderzoek kant. Dat er binnen de techniek ook zo meer nadruk op gelegd kan worden want zover ik weet is dat ook nog niet allemaal rond.

71. Bachelor Student AI, Nijmegen: Als je hele grote informatiecampagnes zou beginnen gericht naar het grote publiek, het volk als het ware, kan je door goed te informeren zou je dan kunnen voorkomen dat er een platform is voor onverantwoorde AI. Waardoor er minder aanleiding is om het op een niet nette manier te doen. Bedrijven die er wél verantwoord mee om gaan zou je subsidies kunnen bieden.

72. Host: Heeft iemand hierbuiten nog een aanvullende oplossing?

73. Bachelor Student AI, Utrecht, 1: Ik denk toch dat de richtlijnen heel erg bijgebracht moeten worden aan iedereen die gaat studeren om daarna iets gerelateerd aan KI te doen. Want ik denk toch wel, je hebt altijd slechte mensen daar kan je niets aan doen, grote bedrijven gaan de boetes gewoon betalen. Maar als ik persoonlijk goede ethische richtlijnen mee zou krijgen tijdens mijn opleiding, zou ik niet snel een baan nemen waarbij alles wat ik heb geleerd meteen het raam uit gaat en dat ik dan iets zit te doen waarvan ik weet dat het slecht is. Ik denk dat het vooral ook uit de opleiding moet komen.

74. Master Student AI, Utrecht, 2: Ik sluit me daarbij aan. Bij het eerst vak van mijn Master hebben we het een week over ethiek gehad, en mijn volledige gedachte daarbij was eigenlijk dat het briljant was dat ze dit nu aan ons meegeven maar ook, waarom in hemelsnaam zit dit niet al in de bachelor? Want ik ken genoeg mensen die na de bachelor KI zeggen, nah mijn startsalaris is prima hoog genoeg, ik ga geen master meer doen. Dus die missen dit volledig, terwijl dit misschien de mensen zijn die per ongeluk iets onhandigs uit kunnen halen. Dus ik denk dat dit in meer vakken, en überhaupt in de bachelor meer besproken moet worden. Dat dit te weinig voorkomt in het curriculum.

75. Master Student AI, Utrecht, 1: Daar sluit me daar enorm bij aan. Ik wil het zelfs wat groter trekken, dat zelfs mensen met slechte intenties het ergens moeten leren. Dus zeker als grote instanties als YouTube en Wikipedia er baat bij hebben, ergens moet je die informatie vandaan halen dus zet het aan het begin van YouTube tutorials en Wikipedia pagina's van machine learning etc., dan moet je er alsnog een keertje langskomen in je educatie of je nou wel of niet ook aan de universiteit hebt gestudeerd. Dat je het gewoon wel overal tegenkomt.

76. Host: De tijd is helaas om, hartelijk bedankt voor jullie komst. Ik zal nu de opname stoppen.

## Appendix B

Focus group invitation:

 ,

Voor mijn bachelor-scriptie ben ik op zoek naar mensen die ervaring hebben met programmeren voor projecten met betrekking tot kunstmatige intelligentie. Het plan is om een gesprek te houden met 6-10 personen. Het zal ongeveer één tot anderhalf uur duren waarin verschillende stellingen/vragen worden bediscussieerd.
Lijkt het je leuk om in gesprek te gaan met mede-programmeurs en heb je woensdag 21 oktober om 19:00 tijd voor een teams meeting?

Stuur dan even een berichtje!

Met vriendelijke groet,

Host Donner
o.j.j.donner@students.uu.nl
06-41549982