Faculty of Behavioral, Management and Social Science

UNIVERSITY OF TWENTE.

The Debate about Responsible Artificial Intelligence in the European Union: Utopian or Dystopian?

A qualitative content analysis of utopian and dystopian visions in the debate about 'Responsible AI' in the European Union between 2016 and 2018

Bachelor Thesis

Due to: July 4, 2018

Name: Philipp Klug

First Supervisor: Dr. Minna van Gerven

Second Supervisor: Dr. Matthias Freise

European Public Administration

University of Twente

Enschede, The Netherlands

Abstract

The age of artificial intelligence (AI) has begun and with it the political debate about the far-reaching moral and ethical implications of the technology. Global political leaders have recognized the enormous potential of AI to benefit our society, but the technology also bears unforeseeable risks. With the goal of promoting beneficial effects of AI while aiming to prevent negative outcomes, a vivid debate amongst institutional actors of the European Union has started to unfold over recent years. Several documents discussing the best regulatory approach to AI were compiled during this process. Within this debate, the notion of a responsible AI emerged and has recently been declared the 'guiding principle of all support for AI-related research' by the European Commission (2018, p. 8). But how does the EU plan to make AI responsible and what is the role of utopian and dystopian visions in this process? This thesis aims at identifying utopian and dystopian elements in the debate about responsible AI by conducting a qualitative content analysis in order to answer the following research question: "Which elements of utopian and dystopian visions are present in the debate about 'Responsible AI' in the European Union between 2016 and 2018?". An answer to the research question will add a better understanding of the impact of utopian and dystopian visions on the debate around responsible AI in the EU. Moreover, the results will show how leading powers in the EU contemplate responsible AI with the goal to preserve - and strive for - a particular set of utopian ideals.

Table of Contents

1. Introduction

Artificial intelligence has arrived. We may not be aware of it, but it is already part of our daily lives. Whenever we make a Google search, algorithms allocate the results for us. Self-driving cars have hit the roads. Whenever we ask Amazon's *Alexa*, Apple's *Siri* or Microsoft's *Cortana*, language-processing algorithms provide us with answers. And looking at Google's latest *Duplex* technology, AI is already able to talk to us with mind-boggling human-like resemblance (Leviathan & Matias, 2018). Bearing such significant advances in artificial intelligence (AI) technology in mind, the prospects of highly sophisticated and even more human-like AI in the near future become increasingly realistic. These new and rapidly evolving technologies have the potential to drastically impact our societies and raise a myriad of moral and ethical questions. The European Union has recognized these potential risks and benefits and a substantial debate around the appropriate regulation of AI in the EU has gained velocity over the recent years. In this debate, the notion of a 'Responsible AI' has emerged and has since become a considerable part of the debate about AI. Moreover, as this research will show, utopian and dystopian visions of robotics and AI play an important role in the formation of views on the (normative) role of AI in our society. From utopian dreams to dystopian nightmares, AI has the potential to bring about both. The answer to the question of AI utopia or dystopia may depend on the degrees of responsibility concerning the actions or inactions of AI itself, which in turn will depend on how scientists and engineers will develop and design AI. Will it be a responsible AI that internalizes sophisticated algorithms corresponding with our moral and ethical codes of behavior? Or will it be an irresponsible AI with unforeseen consequences for our lives, societies and humanity as a whole?

In this context, the debate about AI in the EU has gained significance by the initiative of the Legal Affairs Committee which published an extensive report in January 2017 that discusses specific recommendations to the European Commission on ethical as well as civil-law issues concerning robotics and AI and calls on the European Commission to take legislative action. The report entails particular views on AI and advice to the Commission on how it should be approached and regulated. Prior to the report the European Parliament's Legal Affairs Committee commissioned two documents in relation to ethical and legal aspects of robotics, a study and a briefing. Additionally, three documents by the European Group on Ethics in Science and New Technologies, the European Economic and Social Committee as well as by the European Commission will be part of this content analysis. All six documents play a role in the development of the EU debate on ethical and legal issues of AI and therewith to some extent in the debate about responsible AI. The European Commission has announced to propose comprehensive AI legislation sometime later this year (European Commission, 2017). Hence, it is clear

that the policy debate about AI in the EU has begun. Therefore, the research that will be conducted in this paper will focus on the above-mentioned documents as the corpus that represents this debate to a degree.

1.2 Research Approach

Research in the past focused on the utopian and dystopian visions in the discourse around AI on the platform *Partnership on Artificial Intelligence* (Lamprecht, 2017). However, there is a gap of research focusing on utopian and dystopian visions in political debates, such as the EU debate. No research has yet been conducted involving a 'Responsible AI' concept and content analyses of the EU debate about AI is lacking. Hence, this thesis will adopt an exceptional and explorative approach in that it connects the concept of a 'Responsible AI' with elements of utopian and dystopian visions in order to answer the following main research question:

> "Which elements of utopian and dystopian visions are present in the debate about 'Responsible AI' in the European Union between 2016 and 2018?"

To answer this research question a qualitative content analysis of selected EU documents will be conducted using methods established by Philipp Mayring (2014). The debate about 'Responsible AI' may involve aspects of either moral responsibility or legal responsibility, or perhaps both. Moreover, the connections between 'Responsible AI' and utopian as well as dystopian visions in the data may significantly differ from predetermined theoretical considerations. Therefore, two sub-questions were developed which will help in answering the main research question:

1. Which particular elements of 'Responsible AI' are represented in the debate?
2. How are elements of 'Responsible AI' linked to elements of utopian and dystopian visions in the data?

The theoretical framework will establish the categories according to which excerpts of texts will be categorized regarding the dimensions of the 'Responsible AI concept. Here, an answer to the first sub-question will reveal which of the elements of morally or legally responsible AI are part of the actual debate. The theoretical framework will also establish logical links between the elements of the 'Responsible AI' concept and elements of utopian and dystopian visions. In this case, answers to the second sub-question will clarify how the participants in the debate establish these links.

1.3 Scientific Relevance

Ideally, this research will contribute to the realm of research associated with content analysis in political institutional contexts. Thus, it will contribute to an understanding of how 'Responsible AI' is theoretically constructed in the EU and how it connects to elements of utopian and dystopian visions. The debate on how corresponding officials view and construct a responsible AI is going to be crucial for the formation of policies that are going to affect AI regulation in the future. Through the analysis of how AI and responsibility is discussed in the data, insights will be gained as to how they approach AI responsibility from a political, policy and public administration perspective. AI, especially future versions have a lot of potential to affect our lives in very direct ways. Future AI systems will have to make decisions that will decide over life and death. A near future example of this would be an autonomous car that - for whatever reason - drives into a crowd of people without the possibility to break in time. However, it can still steer the direction, leading to the ethically challenging question of where to steer and whose lives to save of those in the crowd. Possible morally challenging scenarios like these demonstrate that an ethically responsible AI is needed that integrates a deep understanding of our moral values. But how do we decide if an action or inaction by an AI is ethical or not? Scientists and engineers develop AIs through a process called 'reverse engineering' (Miłkowski, 2013). In this process they extract cognitive processes from the human brain, copy so called neural networks and reverse engineer them into artificial networks. Therefore, the question of artificial morality is closely linked to that of human morality and ethics. The foundation of human ethics is meta ethics which certainly must be further discussed by scholars with regards to AI and new emerging research fields such as roboethics or machine ethics. But also other research, such as qualitative content analysis can somewhat contribute to an understanding of how we theoretically and metaphysically construct something like an ethical, responsible AI by conceptualizing it.

Scientists are eventually going to have to somehow translate ethical or responsible behavior into computer code, and AI engineers will have to make sure AI will act correspondingly. Therefore, regulation that enforces ethical and responsible engineering principles can contribute to saver future. But do governments have the necessary knowledge and willingness to make AI responsible today? In the case of this research, the results will demonstrate what role utopian and dystopian visions play in the debate about 'Responsible AI' in the EU. The unmasking of such visions may reveal a biased understanding of how AI works in the minds of institutional actors in the EU, which may lead to misinformed ways of regulation. Moreover, an understanding of how a 'Responsible AI' is approached in the debate by leading EU powers may give an idea about the concerns and objectives of those powers with regards to AI. Will legislation be driven by utopian ideals, or will it be motivated by dystopian fears? Thus, this thesis will

contribute to a better understanding of the role of such visions in political debates about AI in the EU. To what extent contents of the debate will manifest in adopted EU legislation in the future remains to be seen. Therefore, future research may analyze AI legislation in terms of how it reflects responsibility in order for the public to hold governments accountable in case of biased or insufficient legislative action and resulting negative societal consequences.

2. Theoretical Framework

As explained in the research approach, this section will be dedicated to theorizing the main research question. In order to accomplish this, the theoretical framework section will be structured as follows: First, the concept of 'Responsible AI' will be constructed. To construct this concept, it is first necessary to determine the theoretical essence of AI and responsibility. Moreover, different forms of responsibilities and their possible connections to AI must be outlined in order to comprehensively conceptualize 'Responsible AI'. After both, different versions of AI as well as different dimensions of responsibility were demarcated, the concept of 'Responsible AI' will be established. Finally, utopian and dystopian visions will be discussed and specific elements of such visions and their theoretical connection with AI will be elaborated. These elements will later serve as the underlying theoretical foundation for the deductive assignment of categories, which will be necessary in order to identify structures of text that discuss them (Mayring, 2014).

2.1 The Concept of 'Responsible AI'

Throughout the next paragraphs the concept of 'Responsible AI' will be constructed based on theoretical insights concerning different forms of AIs and different dimensions of responsibility. The resulting theoretical conceptualization of 'Responsible AI' is essential for identifying its dimensions and to answer the main research question. Determining the main dimensions of this concept is necessary in order to be able to identify them in the selected data and to link them with elements of utopian and dystopian visions which will be discussed and determined in section 2.2.

2.1.1 What is Artificial Intelligence?

This is a question that is controversial debated amongst numerous academics and outsiders. One of the reasons why is due to the fact that research around AI is very complex as it involves a high number of branches of science such as Cognitive Science, Philosophy, Linguistics, Computer Science, Mathematics and more. Many people would associate AI with human-like robots due to the numerous movies and books that were published in respective genres like science fiction. Such an AI would be called an 'Artificial General Intelligence' (Henceforth: AGI), sometimes also referred to as 'strong AI' (Kurzweil, 2005, p.261). An AGI is defined as being at least as intelligent as a human being, meaning it is able to do any task that a human could do (Nilsson, 1976). However, at this point in time we are not able to create a system that would resemble an AGI. Another notable version of AI is a so called 'narrow AI' or Artificial Narrow Intelligence (ANI). It is often also referred to as 'weak AI' or 'applied AI' (Goertzel, 2013). In this category fall all systems that are able to perform reasoning or problem-solving tasks, such as algorithmic softwares or robotic systems. Thus, all AI systems that are currently in use today fall into the category of an ANI. One last noteworthy classification of AI would be a 'Superintelligence'. According to Nick Bostrom, a superintelligence is defined as 'any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest' (Bostrom, 2002, p.22**)**. Such a superintelligence could emerge from an AGI by continuously improving itself through a concept called 'recursive self-improvement' which would enable it to indefinitely improve itself and its capability to do so resulting in an intelligence explosion and a superintelligence (Kurzweil, 2005; Yampolskiy, 2015).

The link between AGI, superintelligent AI and the concept of a 'Responsible AI' becomes manifest in the 'Control Problem' which presumably will decide over the degree of either utopian, dystopian or in between societal outcomes of the technology. The 'Control Problem' concerns the challenge of how we ensure that such a superintelligent AI will benefit society and not negatively affect or perhaps even subdue or destroy it (Bostrom, 2014). Due to the considerable, yet merely hypothetical existential risk this technology poses to humanity it is worthwhile analysing whether utopian or dystopian visions involving superintelligent AI are present in the EU policy discourse, if any. Hence, AI systems can take one already existing form (ANI) and two hypothetical forms (AGI; Superintelligence). Possible societal threats resulting from these technologies explain the emergence of roboethics and machine ethics along with notions of responsibility in the debate about AI.

2.1.2 Forms of Responsibilities

The question of what exactly responsibility is, is a deeply philosophical one, and the essence of moral responsibility is still being debated in modern times (Eshleman, 2014). The research in this paper does not intend to engage in this discussion, for its focus lies on identifying elements of established forms of responsibility and connecting them first with AI and later with utopian and dystopian visions. In order to identify the theoretical concept of 'Responsible AI' in textual data and to connect it to elements of utopian and dystopian visions, it is important to first make clear what 'responsible' means in this context, what forms of responsibilities exist and how they connect with AI technology. Most notably, two forms of responsibilities have a close relation to the topic of this research and artificial intelligence: Moral responsibility and legal responsibility. Throughout the following paragraphs the two forms will be briefly explained and a possible connection to AI will be theorized based on discussions in relevant academic literature.

Moral responsibility

The question of whether AI is able to bear moral responsibility is subject to a highly controversial debate amongst scholars where some scientists argue that AI cannot be held morally responsible (Friedman & Kahn, 1992; Hew, 2014) while others do indeed foresee the possibility of so called artificial moral agents at some point in the future (Allen et. al, 2010; Matthias, A., 2004). However, the research conducted in this paper does not seek to provide an answer to one of these questions. This research will however attempt to formulate an answer to the following question concerning moral responsibility: If AI could indeed bear moral responsibility, how could such a responsible AI look like? And what makes an AI morally responsible or irresponsible? The discussion of whether AI or future versions thereof can bear moral responsibility concerns fundamental philosophical perspectives on questions of free-will regarding soft and hard determinism, libertarianism or (in-)compatibilism for instance (Ashrafian, 2015). Yet, for the identification of discursive elements concerning 'Responsible AI', definite answers to these questions are not necessary since the analysis in this paper will merely focus on instances of text that supposedly revolve around the theoretical concept of a 'Responsible AI'. While an answer to the question of whether AI can actually be responsible or not will not be answered in this research, the discussions are important to keep in mind due to their unquestionable relevance in the context.

But how could a morally responsible AI look like? What we humans determine to be moral or immoral is a principal concern of ethics. Thus, the actions or inactions of an AI would have to be

analyzed through the lenses of ethical frameworks in order to decide whether the AI acted morally responsible or irresponsible. In his article 'Introduction to the Ethics of New and Emerging Science and Technology' Swierstra (2015) mainly distinguished between consequentialist, deontological (duty), justice and good life ethics, which for the sake of the length of this paper will not be further explained but were extensively discussed by Swierstra (2015). Hence, a morally responsible AI is an AI that would act ethically in the light of the abovementioned ethical frameworks, whereas it can only do this if it was developed, designed, coded or programmed accordingly.

Elements from social responsibility will be included in the construction of the 'Responsible AI' concept. Social responsibility represents an ethical framework that concerns the duty of an entity to act for the benefit of society. Thereby it can be either passive or active: While active social responsibility concerns activities that somehow directly benefit society, passive social responsibility means avoiding activities that could harm society. Hence, a socially responsible AI would be an AI that performs beneficial activities for the society, such as a robot used in elderly care for instance, or one that does not engage in socially harmful acts in the first place. In this context, corporate social responsibility (CSR) has a somewhat meaningful connection to the topic since certain versions of ANI - mostly language-processing algorithms – are already being used for business (see Amazon's *Alexa* or Apple's *Siri* for instance). Thus, CSR in the context of AI means that corporations use AI technology in socially beneficial manners or avoid using AI in socially harmful ways. However, since CSR concerns the way corporations use AI, and not the actions or inactions of AI systems themselves, this form of responsibility will not be included in the 'Responsible AI' concept. In summary, elements of a morally responsible AI primarily concern the adherence if AI to morals, values, ethics, or social responsibility which is only possible if the AI was developed, programmed or designed accordingly.

Legal responsibility

This form of responsibility is a rather practical one as it is directly concerned with legal obligations resulting from region-specific law. Thereby the common terminology in law is 'legal liability' and means 'responsible or answerable in law; legally obligated.' (Black's Law Dictionary, 2014). Under the category of legal liability fall civil as well as criminal law. An example of legal liability is 'product liability' where producers or suppliers of products to the public are responsible in the case of damage caused by these products. A possible connection to AI would be any thinkable product that uses an AI system, like an autonomous car for instance, or a futuristic robot that would cause harm to a human being. In this hypothetical case the producer would be held responsible for the damage, depending on the existing law and corresponding decisions by a court of course. The Committee on Legal Affairs of the European

Parliament initiated the debate on a legal personhood for robotics, which was seen critically by many participating in this political debate. However, in the near foreseeable future, numerous situations are possible which would legally require an AI to act according to legal obligations. For instance, humans are legally obligated to help others who are in need for instance in the EU, thus, a robot could not be held legally accountable in the case of a failure to assist a person in danger, if that robot wouldn't be subject to laws and legal obligations. There are numerous uncertainties around the liability of robotics and AIs in the case of accidents or damage caused by robots. The analysis will show the various viewpoints on these uncertainties of the various actors in the European Union who participate in the debate. Moreover, the analysis will reveal whether certain utopian or dystopian elements are present in the debate about legally responsible AI. Hence, legal responsibility in relation to AI is very much concerned with the legal accountability and legal obligations of AI products or applications and their producers, suppliers or owners. Hence, discussions about a supposed legally responsible AI may contain keywords like liability, accountability or obligation, written in a legal context.

2.1.3 Conceptualizing 'Responsible AI'

By combining the theoretical insights of what kind of systems of AI exist today, or could exist in the future, with the possible forms through which AI could bear moral and legal responsibility, the theoretical construct of 'Responsible AI' emerges.

The theoretical insights gained in section 2.1.1 mainly concerned differentiation between distinctive forms of AI. Thereby forms of AIs that exist today, such as ANIs ('weak AI'), stand in contrast with possible forms AIs could take in the future, such as AGIs ('strong AI') or a superintelligence. This distinction is important to consider when attempting to combine all three AI versions into a single concept, since it not only involves current state of the art technology but also theoretical projections of much more sophisticated, future versions. But for the research conducted in this paper it is necessary to include all three forms in a single concept, since it will not always be possible to know without a doubt whether the author of a certain text refers to ANI, AGI or a superintelligence, if he or she doesn't explicitly uses these or related terms. Therefore, the concept of 'Responsible AI' will involve all forms of possible AIs including weak AI (ANI), strong AI (AGI) and a theoretical superintelligence. As for the forms of responsibilities discussed in 2.1.2 the main distinction was made between the moral- and legal responsibility dimension. Therefore, the analysis of the corpus that will be conducted is going to differentiate between ascribed moral- and legal responsibility. Ideally a 'Responsible AI' should be both, morally as well as legally responsible. If we imagine an AI that talks and interacts with us humans, we would be calmed by the thought that it is doing so in compliance with

ethical principles. But technology may always fail due to defects or other reasons, therefore if an AI is bound by legal responsibilities or obligations, those that would have been damaged by a defect could receive compensation or legal justice, since someone would be legally responsible for the damage by an AI, whether the AI itsself or its developer or owner. However, for this research moral- and legal responsibility will not represent two necessary but two sufficient conditions and facets for and of the 'Responsible AI' concept. In other words, both facets have an *or* relationship, meaning that for this analysis the identification of one of the two facets is a sufficient condition for the concept to enter into force. The analysis may show which dimension of this conceptualization is dominant. Hence, the term 'Responsible AI' will be conceptualized as follows:

A 'Responsible AI' is any system of AI - whether an ANI, AGI or Superintelligence - that through action, inaction bears either moral or legal responsibility.

By means of this conceptualization it will be possible to operationalize this concept by elaborating corresponding dimensions of the two responsibilities in the form of specific keywords through which text about 'Responsible AI' can be identified. These keywords will be derived from the aforementioned theoretical discussions and insights and will be finally determined in the method section of this paper.

2.2 AI in Utopian and Dystopian Visions

This section will explore utopian as well as dystopian visions in relation to AI and technology. Moreover, specific elements that are typical for these visions will be identified in order to connect them later with elements of 'Responsible AI' which were determined in the previous section. By determining elements of either utopian or dystopian visions it will be possible to identify them in the debate about 'Responsible AI' in the corpus, which will be done in the analysis section.

Artificial Intelligence or versions thereof have been subject to numerous books, movies and scientific publications. In many cases AI, or technology in general, was responsible for or contributed to the emergence of utopian or dystopian societies. According to Dinello (2005, p.33) the term utopia was coined by Thomas More in his book *Utopia* (1516) in which social reform, moral example and widespread education would lead utopian perfection. The notion that technology resulting from scientific progress will spiritually enhance and physically liberate humans is central to techno utopians, under the assumption that 'science would understand, control, and perfect nature, including humans' (Dinello,

2005, p.34.). Other aspects of techno-utopian thought concern the revitalization of politics, the protection of human dignity, the protection of other rights such as privacy, impartiality in the sense of nondiscrimination, aspects of social or economic equality, enhanced human freedoms, personal fulfillment and the generation of universal wealth through new technologies (Winner, 1997). In early works of utopian imagination such as in Bacon's *New Atlantis* (1627), the notions of progress, the 'goodness of man' and the positive impact of science and technology are central to techno-utopian future (Dinello, 2005, p.9). Hence, elements of utopian thought in relation to technology and AI mostly include the above-mentioned techno-utopian elements as well as notions of peace, prosperity, liberty (including human freedom and autonomy), harmony, environmental preservation, life enhancement, progress, individual and societal redemption, self-determination as well as control of nature and control of technology (Dinello, 2005). Interestingly, the notion of control is central to both, utopian as well as dystopian visions. While in utopian visions control of nature is crucial to enhance humanity and control of technology is necessary to avoid undesirable, possibly dystopian consequences of technologies, in dystopian visions on the other hand control is typically exerted by actors to maintain political power through oppression or mass-surveillance, typically through the help of technology.

Dystopian visions of societies manifested in literature like *1984* or *Brave New World* in which governments, science and technology work together to control citizens and to enslave them (Dinello, 2005). In the context of AI especially George Orwell's *1984* stands out as it depicts a dystopian, totalitarian society in which computers and machines serve 'Big Brother' by generating cultural products in order to suppress new ideas, and by providing assistance for propaganda, control, torture and surveillance (Dinello, 2005). The movie *Alphaville* depicts an extreme version of a supercomputer which eliminates human values and exercises absolute control and total surveillance (Dinello, 2005). A variety of utopian and dystopian visions involving such a superintelligence exist. Max Tegmark categorized such possible scenarios ranging from utopian visions where superintelligent AI and human coexist peacefully to dystopian ones where such an AI would either control or destroy humanity entirely (Tegmark, 2017), which leads to the establishment of the dystopian elements of losing control over technology and subjugation. Moreover, dystopian visions involving various (possible future) versions of AI typically concern scenarios in which either totalitarian governments make use of AI to control and surveil a society sometimes accompanied with social inequalities, or scenarios in which an advanced AI itself represents the instance of ruling power and societal control. In these cases, AI is either used irresponsible by actors or authorities to achieve their (totalitarian) ends, or AI itself is acting irresponsibly through oppressive and inhuman actions. Many dystopian visions also depict scenarios where those who are in possession of powerful technology - typically multinational corporations – reap its benefits while the situation for rest of the population is deteriorates (Dinello, 2005). Therefore, dystopian visions in relation to technology

and AI may include words like war, chaos, poverty, suppression, oppression, control of power, domination, surveillance, slavery and inequality.

The theoretical connections between responsible or irresponsible AI and the elements of the visions vary significantly in their plausibility. Whereas the link between responsible AI behavior and the protection of human dignity is rather clear, other connections such as the impact of irresponsible AIs on wars leaves more room for possible scenarios. The currently often discussed lethal autonomous weapons (LAWs) could theoretically be able to violate international law in certain warlike scenarios (Lin et. Al 2009). For example, a responsible LAW (if a lethal AI could ever be described as morally responsible) would not endanger civilians and only focus on actual threats such as terrorists. On the other hand, this is never guaranteed one does not want to imagine the consequences of a LAW that does not comply with ethical principles. These consequences would indeed be dystopian, leading to possible scenarios where unmanned, autonomous drones fly over villages and kill every single human being in a certain conflict. Even today remotely operated drones bomb villages in the middle-east in the fight against terrorism. They are still operated by humans who have to make the decision whether to engage or not, therefore a near future scenario where autonomous drones make this decision by themselves according to the information they have gathered through sensors is certainly not impossible (Arkin, 2010). Hence, irresponsible AIs can without a doubt spark, contribute or lead to numerous dystopian realities. However, the link between irresponsible AI and dystopian elements such as war, suppression, oppression, control of power, losing control over technology, domination, discrimination, slavery and surveillance is much clearer than the connection to poverty and inequality. The latter two elements describe dystopian economic visions which are rather the result of the existence of AI itself and the way governments or rules organize society, than the irresponsible actions or inactions by an AI itself.

The same logic applies to the connection of utopian elements and responsible AI. Hereby a responsible AI can clearly contribute to the protection of human rights, by for example reporting violations to authorities. Moreover, a responsible AI can protect and enhance human liberty, by firstly not restricting it through irresponsible behavior, but also enhancing their freedom by providing socially disadvantaged people such as people with handicaps or old people with physical limitations for instance with support. On the other hand, it is implausible to assert that the generation of universal wealth would depend on responsible or irresponsible behavior by an AI, as this would only be possible in very indirect ways. As with the aforementioned dystopian economic elements, the fact that AI can surely contribute to more wealthy and prosperous society has intrinsically nothing to do with aspects of responsibility of the AI itself but rather with the socio-economic responsibilities of governments to ensure a fair and just distribution of the economic benefits that AI will bring about. Nevertheless, a responsible AI can certainly contribute to a more equal society, not in economic terms, but in terms of the way people of different

ethnicities, genders or backgrounds are being treated. Here a current example would be a facial recognition software that does not discriminate people in the process of job applications. Research has shown that an AI could act discriminatory if the data that has been used to develop the AI was incomplete or biased (Bryson et al., 2017). Hence, to prevent AI from acting irresponsible, and to prevent dystopian visions from becoming a reality, it is important that scientists and engineers create AI in such a way that risks of irresponsible behavior is minimized and that AI behaves responsibly by design.

2.3 Concluding remarks

By means of the theoretical conceptualization of a 'Responsible AI' it will be possible to identify concrete elements in instances of texts that concern AI in the context of either moral responsibility or legal responsibility. In addition, through the determination of concrete theoretical elements of dystopian and utopian visions it will be possible to identify them in structures of text.

3. Method

The method section has the following structure: First, the general method of this research will be presented and justified. Second, it will be explained what cases were selected for analysis and why, briefly introducing and discussing the selected documents and their relevance within the overarching discursive context. Consequently, the method of how the data was collected will be outlined. Finally, the operationalization of the data will be presented and justified with regards to theoretical findings and inductive category assignment.

3.1 Qualitative Content Analysis according to Mayring

The method that will be used in this research is a Qualitative Content Analysis (Henceforth: QCA) according to Philipp Mayring (2014). For the objective of the research of this thesis, Mayring's QCA approach was deemed especially viable due to its systematic yet adaptive method of content analysis, which will be further discussed and accounted for in the paragraphs below. According to Steigleder (2008) QCA 'has proven its worth in many studies. With its different techniques of analysis and its methodological concept it is excellently adapted to analyze qualitatively collected material' (p. 197). The strength of Mayring's approach to QCA lies in its systematic, step by step procedural model of research. He suggests several distinctive stages of qualitative research methods that have to be followed in order to

arrive at reliable and valid results. This research in this paper will closely follow this procedural model (an illustration of this model can be found in appendix C). However, Mayring (2014) points out that QCA 'is not a standardized instrument that always remains the same', but has to be adapted for the specific objects, materials and issues that are being researched (p. 39). Therefore, we have to adapt Mayring's QCA to this research in order to arrive at systematic, yet distinctive research design.

This research design primarily depends on the type of data and the goal of this research which is to elaborate an answer to the initial research question. The type of data is textual data, therefore the coding unit, which represents the smallest component of the data that can be analyzed, is one sentence, since single words are not enough to represent a debate about 'Responsible AI' that involves utopian or dystopian elements. The context units, which is the largest possible component that can be analyzed according to Mayring (2014) are in this case groups of sentences in the form of excerpts. The recording unit consists of all six documents, whereas different categories can be assigned to different coding units within these documents. Determining these units is important to ensure intersubjectivity, so that different analysts can comprehend and reproduce the analysis (Mayring, 2014). Concerning the procedure of analysis and interpretation that will be followed in order to formulate an answer to the research questions by extracting and revealing (hidden) textual structures and elements in the data, the method of Deductive Category Assignment (Structuring), which according to Mayring (2014) represents a form of interpretation that aims at filtering out specific aspects of textual data, depending on pre-established criteria, will be adopted. Another argument for this method is the fact that the main research question is descriptive since it doesn't ask for an explanation, but for a description of which elements are present in the data. In such a case, Mayring (2014) suggests a descriptive research design that applies a 'deductively formulated category system' to identify the specific occurrence of those categories in a text (p. 12).

For this research it means that elements of the 'Responsible AI' concept as well as elements of utopian and dystopian visions have to be systematically categorized, in order to identify them in the documents. Hereby, 'deductively formulated' means that the categories will be developed according to the elaborated theoretical findings and considerations in the theory section. According to Kohlbacher (2006) the system of categories is 'the core and central tool of any content analysis' (p. 58). Therefore, the coding of the units of analysis - which means their allocation to a particular category - has to be done with great care and reasonable diligence in order to ensure intersubjectivity of the procedures and reproducibility of the results of this research. The operationalization and categorization will be discussed in the section 3.4. After the category systems and the coding guidelines were defined, Mayring (2014) states that a preliminary material run-through should be conducted in order to test whether the categories are applicable, and whether definitions and encoding rules lead to feasible categorical allocation. During this trial run-through the so called 'points of discovery' should be marked, copied out of the text and

assigned to a specific category (Mayring, 2014). Consequently, possible ambiguous categorization of results will eventually lead to a revision of the definitions of the category system, after which the main run-through will be carried out where once again points of discovery will be extracted and processed (Mayring, 2014). The extracted data in the form of excerpts will then be thoroughly analyzed in in order to answer the main- and sub research question, culminating in the conclusion. To ensure reliability of the results Mayring (2014) suggests a *Re-test* in which the research procedure is carried out once more to see whether it will lead to the same results. Moreover, to ensure construct validity the operationalization will be evaluated in terms of its adequacy in relation to theoretical underpinnings, and the final results will be assessed according to their plausibility bearing in mind theoretical considerations and expectations (Mayring, 2014).

3.2 Case selection

The case of this research is the debate about responsible AI in the EU. While legislation on AI with respect to issues of moral responsibility or legal liability has neither been adopted or proposed on the EU-level at this point in time, the political debate focusing specifically on (responsible) AI and associated topics which at some point in the future will eventually lead to EU regulation or directives has indeed begun to unfold in recent years. The cases were selected according to their significance in the time frame, to their relevance in the political debate on AI as well as according to their meaningfulness in the context of AI policy debate in the EU. Thus, the cases that were selected for this research represent the early stages of policy debate around responsible AI in the EU.

The first case that was selected for this research is the 'Report with Recommendations to the Commission on Civil Law Rules on Robotics'. The report was published on January 27, 2017 and its rapporteur was vice-chairman of the European Parliament's Committee on Legal Affairs Mady Delvaux. The document represents a request to the European Commission 'to submit a proposal for a directive on civil law rules on robotics' based on Article 225 TFEU and Article 114 TFEU (European Parliament, 2017). The report has a total length of 64 pages and includes numerous recommendations concerning general principles around legal, ethical and other societally relevant topics around robotics and AI. In addition, it includes the opinions of seven committees of the European Parliament. The second selected case is a study on the future of 'Civil Law Rules on Robotics' published on October 12, 2016 by the European Parliament's Policy Department for Citizens Rights and Constitutional Affairs after it was requested by the European Parliament's Committee on Legal Affairs. It has a total length of 34 pages and the topics in the study range from general considerations on robotics and resulting liability issues to an

analysis of ethical principles in the development of robotics. The third selected case is a briefing on the 'Potential Benefits and Ethical Considerations' of Artificial Intelligence by IBM research scientist Francesca Rossi for the European Parliament, also published on October 12, 2016. The document is eight pages long and includes general discussions on AI in relation to topics like computing power, data, ethics and trust as well as AI and policies whereby the role of IBM is given special attention in the context.

The fourth selected document named 'Artificial intelligence - The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society' was written by the European Economic and Social Committee. It has a length of 13 pages and was published on May 31, 2017. It represents the committee's opinion on AI and was published on the committee's own initiative according to Rule 29(2) of the Rules of Procedure. The fifth document "Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems" represents an advisory report by the European Group on Ethics in Science and New Technologies and is directed towards the president of the European Commission. It is 24 pages long, was published on March 9, 2018 and includes numerous statements in relation to ethical considerations of robotics and AI. The sixth and final document 'Artificial Intelligence for Europe' represents a communication from the European Commission to various actors in the EU. It has a total length of 20 pages and was published on April 25, 2018. It has a special role in the EU debate about AI since the European Commission will is the only institutional body of the EU that can propose legislation. Therefore, the views and opinions expressed in this document are likely to be a part of actual, future legislation. In response to the motion for a resolution initiated by the Legal Affairs Committee, the European Commission has declared its intention to propose legislation on AI later this year (European Commission, 2017).

3.3 Data Collection

All documents that were collected are official, publicly available documents by the EU and its institutions, made available on the official websites of the EU. The criteria according to which the data was collected is their relevance in the context of the EU debate about responsible AI. The data was published in the time frame between 2016 and 2018, and therefore concerns a highly up-to-date topic. The 'Report with Recommendations to the Commission on Civil Law Rules on Robotics' was published on the initiative of the Committee on Legal Affairs in 2017, which also commissioned the study which was published on the exact date as the briefing in 2016. Therefore, the study as well as the briefing document are closely related to the creation of the report, due to their evident connection to the Legal Affairs committee. Therefore, all data is interconnected within the debate about responsible AI. As stated earlier, in response to the request in the report to propose legislation, the European Commission will propose

corresponding regulation at some point later this year (European Commission, 2017). Therefore, the report plays a major role in the EU policy process and debate about AI, that will eventually lead to EU regulations or directives. Due to the fact that the study was requested by the Committee on Legal Affairs it is closely connected to the development of the first selected case and is therefore an integral part of the policy debate about AI in the EU. Unlike the other documents which were written by EU officials, the briefing document is the work of an individual who worked for 'The International Business Machines Corporation' (IBM) at the time of the creation of the document. Therefore, due to the resulting potential conflict of interest, special attention will be given to the institutional and corporate power related context of the document in the analysis, which might be important for the revelation of utopian or dystopian elements. Nevertheless, due to the fact that this briefing was published on the same date and by the same policy department as the second selected case and was intended to brief members of the European Parliament and its committees on the topic of AI, it plays considerable and special role in the policy debate about AI in the EU due to organizational position of the author. The document by the European Economic and Social committee will predominantly discuss AI in relation to economic terms. Therefore, as argued in the theory section, it is expected that rather few excerpts will concern utopian or dystopian elements, due to the insignificant role of AI responsibility in the realm of economics. On the other hand, the document by the European Group on Ethics in Science and New Technologies is expected to discuss 'Responsible AI' a lot, due to the obviously strong relationship between ethics and responsibility. As for the communication by the European Commission, the analysis will probably reveal the most significant insights concerning actual legislation as the commission is the only institutional body of the EU that can propose legislation for adoption, while it will be up the European Parliament and – depending on the concerned policy areas – the Council of Ministers to adopt it or not.

3.4 Operationalization & Categorization

In order to be able to conduct the analysis it is necessary to operationalize the research question into categories through deductive category assignment (Mayring, 2014). For this research it means that the elements of utopian and dystopian visions as well as of morally and legally 'Responsible AI' have to be operationalized and categorized according to the pre-established theoretical elements.
The 'Responsible AI' concept will consists of two sufficient facets or dimensions, namely moral responsibility and legal responsibility. These facets of 'Responsible AI' have an *or* relationship to the concept, meaning that for this research the identification of one of the dimensions in a text is sufficient for the text to concern the debate about 'Responsible AI. Both morally as well as legally responsible AI can be identified in the text through specific keywords. These keywords were derived from the theoretical

insights in the theory section following the deductive category assignment model of Mayring (2014). These coding guidelines will be used in order to identify structures of text in order to be able to adequately categorize them either into the category of responsible AI or into the category of utopian and dystopian visions. For the process of analysis, several category tables were developed containing four columns of information after Mayring's (2014) methodology: Category label, category definition, anchor example and coding rules. Hereby anchor examples serve as exemplifications for structures of text that belong to a particular category. An example of such a table can be seen below in Table 1. The entirety of the category tables can be found in Appendix A.

Dimension	**Anchor example**	**Coding rule**
Subcategory R1: Morally responsible AI	'whereas the Union could play an essential role in establishing basic **ethical principles** to be respected **in the development, programming** and use **of robots and AI** (…)' (EU4, 2017, p. 3)	Ethical/value principles/framework for the development/programming/design of AI

Table 1

This table depicts the subcategory of the 'Responsible AI' concept. More accurately, it shows the 'morally responsible AI' dimension, an anchor example from the corpus, and a small component of the numerous coding rules through which text that discusses this dimension can be identified. In this case, the anchor example represents a discussion about morally responsible AI due to the application of the visible coding rule in combination with the direct use of corresponding words marked in bold.

However, sometimes, the authors of the documents will not explicitly mention the keywords listed in the coding rules column and will instead perhaps only paraphrased or describe social responsibility for instance. Thus, in some instances of the analysis it will be necessary to interpret the elements within the semantic and thematic context of the discussions. Due to interpretation leading to threats regarding the reliability of results, the research in this paper will abstain from quantification of results. However, the analysis may show if certain elements of 'Responsible AI' or utopian and dystopian visions are clearly dominating the debate, therefore cautious estimations concerning the extent to which elements of morally or legally responsible AI as well as elements of utopian and dystopian visions respectively dominate will be made in the conclusion. According to Coffey & Aktinson (1996) 'coding can be thought about as a way of relating our data to our ideas about these data (p.27). Therefore, the coding rules, according to which excerpts will be categorized, represent the explorative dimension of the research conducted in this paper by attempting to relate the conceptualization of a 'Responsible AI' - which resulted from inductive reasoning - with the specific textual data of the corpus. By means of the coding rules and keywords the data will be categorized and analyzed according to the QCA method.

3.5 Concluding remarks

In this section, the QCA approach after Mayring (2014) was outlined and justified. Moreover, the case selection and data collection methods were explained. The operationalization and categorization of the 'Responsible AI' concept as well as of the elements of utopian and dystopian visions is closely related to the theoretical insights gained in section 2.1 and 2.2 in this paper and is therefore deductive. The resulting coding rules combined with the QCA approach will be applied in the analysis in order to answer the main and sub research questions.

4. Analysis

This section is dedicated for the analysis of the selected corpus. For reasons of convenience and readability, each document was coded will be analyzed in a sequential order. The codes for the documents can be found in appendix B. Throughout the analysis, it will be clearly indicated why certain excerpts and parts thereof can be allocated to a particular category. Moreover, it will be explained which elements of utopian and dystopian visions can be found in these excerpts in order provide an answer the research question in the conclusion. As argued in the method section, in some instances certain categorization is only possible through interpretation of the thematic or socio-cultural context in which the texts were written.

4.1 Report with recommendations to the Commission on Civil Law Rules on Robotics

This report represents the motion for a resolution and is aimed at the European Commission to propose legislation concerning civil law rules on robotics. With 64 pages it is the longest document of the corpus and includes text written by a number of representatives of the various committees, which will be clearly indicated in order to appropriately assign the institutional backgrounds of the authors, which may add insights to the context of excerpts.

> 'whereas the Union could play an essential role in establishing basic ethical principles to be respected in the development, programming and use of robots and AI and in the incorporation of such principles into Union regulations

> and codes of conduct, with the aim of shaping the technological revolution so that it serves humanity and so that the benefits of advanced robotics and AI are broadly shared (...)' (EU3, 2017, p. 6)

This text is an excerpt from the general principles written by the Committee on Legal Affairs. Through the application of the coding framework we can say with certainty that this text belongs into subcategory R1 (morally responsible AI), since it calls for 'establishing basic ethical principles (...) in the development, programming and use of robots and AI' (EU3, 2017, p. 6). As argued in the theory section, there can be no responsible AI without it being designed and programmed to follow ethical principles. In the same sentence, the goal is described as 'shaping the technological revolution (...) so that the benefits of (...) AI are broadly shared (...)'. Through interpretation we can identify two elements of utopian visions here. First, one cannot shape a technological revolution without controlling the technology itself. Therefore, the utopian aspect of control of technology is clearly present, also because the term 'controlling' could replace 'shaping' without necessarily changing the meaning of the sentence significantly. Moreover, the sentence 'so that the benefits of (...) AI are broadly shared' (EU3, 2017, p. 6) clearly describes the egalitarian principle and utopian element of *equality*, since the benefits of the technology should be shared.

> 'Points out that the guiding ethical framework should be based on the principles of beneficence, non-maleficence, autonomy and justice, on the principles and values enshrined in Article 2 of the Treaty on European Union and in the Charter of Fundamental Rights, such as human dignity, equality, justice and equity, nondiscrimination, informed consent, private and family life and data protection, as well as on other underlying principles and values of the Union law, such as non-stigmatisation, transparency, autonomy, individual responsibility and social responsibility, and on existing ethical practices and codes;' (EU3, 2017, p. 10)

In this excerpt the identification of the 'Responsible AI' category results from the textual context. The overarching headline of this excerpt is 'General principles concerning the development of robotics and artificial intelligence for civil use' (EU3, 2017, p. 8). Therefore, the content of the excerpt, namely 'the guiding ethical framework should be based on (...)' belongs to category R1, since the guiding ethical framework relates to ethical the development which is part of the coding rules of the category. Thus, the text revolves around a framework to make AI ethical, which is a sufficient condition for the 'morally responsible AI' category as part of the operationalized 'Responsible AI' concept.

Within this excerpt we can identify a number of utopian elements. Analyzing the text we can identify the utopian elements of *liberty* (*autonomy*), *justice*, *human dignity*, *equality* and *non-discrimination*. Through interpretation we can also find the utopian element of *privacy* since data-protection is typically intended to ensure the privacy of personal data.

> ‘License for designers: You should take into account the European values of dignity, autonomy and self-determination, freedom and justice before, during and after the process of design, development and delivery of such technologies including the need not to harm, injure, deceive or exploit (vulnerable) users.’ (EU3, 2017, p. 25)

In this case we can identify aspects of social responsibility through interpretation, regarding the use of the words ‘not to harm, injure, deceive or exploit (vulnerable) users.’ (EU3, 2017, p. 25) since a socially responsible AI acts for the benefit of society, obviously without harming, injuring or exploiting people. Thus, the text clearly not only asks the designers to be responsible but also to design AIs to be socially responsible, due to the use of the words ‘during and after the process of design, development and delivery of such technologies’ (EU3, 2017, p.25). Hence, this phrase can be allocated to the R1 subcategory. Utopian values of *dignity*, *liberty* (*autonomy*), *self-determination* and *justice* can be clearly identified as well by analyzing the use of keywords.

> ‘The JURI Committee believes that the risks posed by these new interactions should be tackled urgently, ensuring that a set of core fundamental values is translated into every stage of contact between robots, AI and humans. In this process, special emphasis should be given to human safety, privacy, integrity, dignity and autonomy.’ (EU3, 2017, p. 27)

Through interpretation we can conclude that the phrasing ‘ensuring that a set of core fundamental values is translated into every stage of contact between robots, AI and humans.’ (EU3, 2017, p.25) concerns the debate about morally responsible communication between AI systems and humans. Consequently, it is clear that during this contact between AIs and humans, the utopian elements of *human safety*, *privacy*, *dignity* and *liberty* (*autonomy*) should play a role due to the use of the words ‘In this process, special emphasis should be given to (...)’ (EU3, 2017, p.25). Hence, according to this excerpt, contact between AIs and humans should be based on fundamental values which, except for integrity, also resemble elements of utopian visions.

> ‘Believes that robotics and artificial intelligence, especially those with built-in autonomy, including the capability to independently extract, collect and share sensitive information with various stakeholders, and the possibility of self-learning or even evolving to selfmodify, should be subject to robust conceptual laws or principles, such as that a robot may not kill or harm a human being and that it must obey and be controlled by a human being;’ (EU3, 2017, p. 36).

Contrary to the excerpts before, this text was written by the Committee on Civil Liberties, Justice and Home Affairs. Through looking at the text we can conclude that the committee calls for a set of laws and principles that should be an integral part of AI due to the phrasing ‘should be subject to (...)’ (EU3, 2017,

p. 36). We may also interpret this statement as calling for a legally responsible AI that is subject to laws, hence the excerpt can be allocated to the R2 category. Moreover, by interpreting the sentence '(...) that a robot may not kill or harm a human being and that it must obey and be controlled by a human being;' (EU3, 2017, p. 36) we can state with certainty that it concerns a socially responsible AI, that does not kill or harm a human being, in combination with the utopian element of *control* over the technology, since the AI necessarily has to obey and be controlled by a human being.

However, there is a deeper meaning within this excerpt. We may expand this analysis by interpretation and while taking into account the socio-cultural context. Through interpretation in context we can derive information from the fact that, as mentioned before, this text was written by the Committee on Civil Liberties, Justice and Home Affairs. Bearing this in mind, we can add another layer to the text, namely the aspect of utopian *liberty* resulting from the utopian element of *control*. Since this committee is concerned with civil liberties, it seeks to protect the liberties of people. An autonomous AI that does not obey human command and that is not controlled by humans can seriously endanger human liberty through its actions. This issue becomes clearer if we add the contemporary socio-cultural context, namely the societal conditions in which this text was produced, more precisely, the situational time aspect including the lack of scientific knowledge around AI. Relating back to the 'control problem' in the theory section, there is no scientific clarity on how we would be able to ensure that sophisticated versions of AI act according to the volition and benefit of humanity. In relation to the afore cited text, the phrasing

> '(...) that artificial intelligence, especially those with built-in autonomy, including the capability to independently extract, collect and share sensitive information with various stakeholders, and the possibility of self-learning or even evolving to selfmodify, should be subject to (...) (EU3, 2017, p. 36)

clearly acknowledges the possibility that AI could self-modify. Once again applying the third dimension, the term 'self-modification' has a powerful sociocultural meaning with regards to AI, as it describes a situation in which an AI modifies its own algorithms. However, those algorithms decide over whether the AI acts for the benefit of humanity, over whether it protects or harms a human, and over whether it intrinsically behaves responsible or irresponsible. Additionally, the use of the word 'evolving' shows that the Committee is well aware of the fact that AI, once it has reached a certain level of intelligence, is able to teach itself through machine learning to increase its own intelligence and intellectual capability, including the ability itself to learn. A process which has been discussed in the theory section that could ultimately lead to a superintelligence, which represents an existential threat to humanity according to experts in the field, as discussed before (Bostrom, 2014)**.** Hence, by not only interpreting but also considering the socio-cultural context when analyzing this excerpt, we reveal a far-reaching set of attached meanings, fears and ideologies. First the committee calls for a legally and morally responsible AI

that follows a set of laws and (ethical) principles, second they discuss the possibility of AI to evolve and self-modify, while third calling for humans to hold the ultimate utopian control and upper hand throughout these developments. An utopian ideal of control, which is with regards to the current sociocultural context in which the excerpt and this analysis is written regarding our current scientific knowledge of AI technology, including powerful, possible future versions of AI, by no means guaranteed.

Following the discussion around a set of ethical principles that shall guide the development and design of AI, the committee continues with a number of points of discussion concerning privacy and data-protection aspects in this process. In the context of ethical, thus morally responsible AI, the committee highlights 'the importance of preventing mass-surveillance through robotics and artificial intelligence technologies;' (EU3, 2017, p. 37). Through the application of the coding framework we are able to identify the dystopian element of *mass-surveillance* in the text. Hereby an important aspect of intertextuality between the headline 'Privacy and data-protection' (EU3, 2017, p.37) - whereas *privacy* represents an utopian element - and the dystopian element of *mass-surveillance* can be identified through contextual interpretation and societal explanation. In the societal context, there can be no privacy and data-protection along with mass-surveillance and vice versa. The dystopian fear is fueled by the algorithmic capabilities of AI to collect and categorize personal data, through various possible modern and future sources such as computers, smartphones or in the future sensors and cameras of robots. Given the hypothetical yet not impossible situation in which a huge part of the population owns a personal robot in their homes who is connected to the internet, the potential and feasibility of mass-surveillance through whichever actor is a plausible scenario. In the context of 'Responsible AI' this could mean that an irresponsible AI would share private data, while a responsible one would protect it and only share it with the explicit consent of the person. Hence, in this case contextual interpretation reveals that the opinion of the committee is characterized by dystopian fears of *mass-surveillance*, including the various possible ways through which AI could exercise or contribute to it, while the committee strives for the utopian ideal of *privacy*, whereas data-protection cannot be guaranteed without an AI that gathers and handles data responsibly.

> 'Research Ethics Committees (RECs) should take into account the ethical questions raised by the development of medical robotic devices and CPS in many areas of healthcare and assistance provision to disabled and elderly people. Issues such as equality of access to robotic preventive health care (...) should be given due consideration. RECs and the Commission are encouraged to start a reflection in order to develop a code of conduct for researchers/designers and users of medical CPS, that should be based on the principles enshrined in the Union's Charter of Fundamental Rights (such as human dignity and human rights, equality, justice and equity, benefit and harm, dignity, non-discrimination and non-stigmatisation, autonomy and individual responsibility, informed consent, privacy and social responsibility as well as the rights of the elderly, the integration of persons with disabilities, the right to healthcare, and the right to

> consumer protection) and on existing ethical practices and codes. It is noteworthy that robotics can introduce a high level of uncertainty regarding responsibility and liability issues.' (EU3, 2017, p. 52-53)

This excerpt was written by the Committee on the Environment, Public Health and Food Safety. It represents the committee's opinion on the use of robotic devices and so called cyber-physical-systems (CPS) which are for example algorithmic monitoring systems. By applying the coding-rules on the text we can conclude that it concerns 'a code of conduct for researchers/designers and users of medical CPS' (EU3, 2017, p. 53) based on the EU Charter of Fundamental Rights which includes many elements of utopian visions such as *human dignity*, *equality*, *justice*, *impartiality* (*non-discrimination*), liberty (*autonomy*) and *privacy*. Since the author explicitly includes researchers and designers of CPS to those who should follow the code of conduct, which also includes social responsibility, this excerpt is part of the morally responsible AI category R1, since - as it was argued in the operationalization - there can be no morally responsible AI without scientists, designers and engineers deliberately developing it according to moral and ethical values. Going a step further and taking into account the socio-cultural context, we may add another in-depth layer to the analysis. The sentence 'Issues such as equality of access to robotic preventive health care (...) should be given due consideration.' (EU3, 2017, p. 52) reveals a deeper aspect of equality and has to be explained taking into account our socio-cultural practices concerning the currently existing healthcare systems. While in the EU the right to healthcare is guaranteed, in other western societies the question of whether you receive your treatment or not depends whether you are actually insured or can financially afford it. With regards to socio-economic inequalities it would indeed be a dystopian nightmare if a healthcare robot denies you a lifesaving treatment because you do not have insurance or enough money. In such a dystopian society, only the rich would be able to receive healthcare, while the poor would not. Therefore, in this context the dystopian element of *inequality* can be identified. Going back to the aforementioned utopian elements acquired through the textual analysis and adding another layer to the utopian element of equality, we are able to determine the full picture of utopian ideology in the minds of the committee concerning AI in healthcare. In the view of the committee, your dignity, liberty and privacy should be respected by (advanced) robotics and AIs. Further, these systems should not discriminate against you, respect your autonomy and as a human being, your human dignity should be respected by providing you with equal access and a right to healthcare. Hence, it is quite difficult to imagine a more utopian picture of AI in the context of healthcare.

4.2 Artificial Intelligence: Potential Benefits and Ethical Considerations

This document was written by IBM research scientist Francesca Rossi and is eight pages long. It represents a briefing for the members of the European Parliament. Due to the affiliation of the author being an employee at a company that makes profits through AI systems, it is expected that regulation of the technology, which could potentially reduce profits of the company, will be viewed critically, while the capabilities and possibilities of the technology will be praised.

> 'But trust will also require a system of best practices that can help guide the safe and ethical management of AI systems including alignment with social norms and values; algorithmic responsibility; compliance with existing legislation and policy; assurance of the integrity of the data, algorithms and systems; and protection of privacy and personal information.' (EU2, 2016, p. 4).

Applying the coding rules it is clear that the text belongs to the morally responsible AI category R1 due to the use of the term 'algorithmic responsibility' (EU2, 2016, p. 4). As discussed in section 2.1.1, algorithms represent ANI, and therefore a form of AI. Thus, the coding rule 'AI responsibility' applies. When talking about the 'safe and ethical management of AI systems including alignment with social norms and values' (EU2, 2016, p.4) the author does not explicitly name which norms and values are meant making it impossible to identify possible utopian or dystopian elements which could be have been derived from specific values, as earlier examples in this analysis have shown. However, the author does mention 'protection of privacy and personal information' (EU2, 2016, p. 4) whereas *privacy* is an element of utopian ideology, as discussed before.

> 'One of the primary reasons for including algorithmic accountability in any AI system is to manage the potential for bias in the decision-making process. This is an important and valid concern among those familiar with AI. Bias can be introduced both in the data sets that are used to train an AI system, and by the algorithms that process that data. At IBM, we believe that the biases of AI systems can not only be managed, but also that AI systems themselves can help eliminate many of the biases that already exist in human decision-making models today.' (EU2, 2016, p. 4).

Analyzing the text, this excerpt can be put in the legally responsible AI category R2, as it uses the term 'algorithmic accountability' (EU2, 2016, p. 4). Once again, algorithms are a form of weak AI, therefore the coding rule 'AI accountability' applies. In this context, the author discusses bias in the decision-making process of an AI, which reveals the dystopian element of *discrimination* through interpretation. Consequently, the author expresses the view that at IBM - the company where she is employed - they believe that AI can assist in combating existing human bias, which can be interpreted as following the

utopian ideal of *impartiality* or *non-discrimination*. Hereby she does not explain how AI could assist in decreasing human bias, which puts the phrasing in the light of exaggerated technology glorification. Human fallibility is a part of human nature, therefore the belief that by means technology we can control human nature and thus eliminate human fallibility such as human bias, is central in technological utopianism. Hence the utopian element of *control of nature* is somewhat visible here.

> 'Ethical issues, including safety constraints, are essential in this respect, since an AI system that behaves according to our ethical principles and moral values would allow humans to interact with it in a safe and meaningful way.' (EU2, 2016, p. 7)

This is another example where the excerpt clearly belongs to the R1 category due to the sentence '(...) an AI system that behaves according to our ethical principles and moral values (...)' (EU2, 2016, p. 7), however, once more she does not name which values or ethical principles she has in mind. It appears that, for some reason, she does not want to engage in the discussion around the specific moral values that should play a role in AI. The author does however name the utopian element of *safety*, which according to her should result from a responsible AI.

> 'It is clear that a lack of regulations would open the way to unsafe developments. However, also excessive regulations would have a cost to society, since they would not allow us to take advantage of all the potential benefits that AI can bring, such as saving lives, curing diseases, and solving planetary problems.' (EU2, 2016, p. 7)

This excerpt follows directly from the afore cited text and can therefore fits into the R1 and V1 category with regards to the utopian element of *safety*. The author states that too little regulation could lead to unsafe developments, but in the next sentence argues that too much regulation could have negative effects as well by the use of the words 'cost to society' (EU2, 2016, p. 7). This is a quite remarkable linguistic composition as it suggests that society would have to face costs or in other words, society would lose something. This is a cunning yet subtle proposition as the formulation leads to the factually wrong conclusion that too much regulation will invoke costs for society. Gaining less is not the same as losing something: While it is true that excessive regulations makes innovation in AI more difficult and using the author's words 'they would not allow us to take advantage of all the potential benefits that AI can bring' (EU2, 2016, p. 7) saying that one will have less benefits is not the same as saying that one will lose something, by having a to pay the cost. While opposing too much regulation is certainly not unreasonable, it is no surprising that the author is critical of regulation. Here, the authors institutional affiliation confirms what has already been suspected earlier. Political debate around the regulation of technology, especially if it involves an affiliate of a private, profit seeking organization, is also always a power play.

Public interest and private interest are not congruent, and when it comes to a promising technology like AI, the (profit) stakes are high. Thus, considering the socio-political context, we may interpret the dystopian element of *control of power* in this case. In techno-dystopian visions, power is not only exerted by totalitarian governments but also by corporations. Corporate profit is driven by technological progress which in turn depends is facilitated by those who control the market of the technology. Therefore, corporations have an enormous interest in preventing profit-threatening regulation of technology and they do this by influencing and controlling those political powers who are able to adopt regulations. In dystopian visions corporations 'strongly influence the political processes that ostensibly regulate technology's development (…) demonstrating that it mirrors the corruption of corporate manipulation' (Dinello, 2005, pp. 273-274). In relation to responsible AI this interpretation could mean that multinational corporations such as IBM have few interests in the development of a 'Responsible AI' since this would imply regulation of the technology which could reduce their profits and power.

In summary, while the author does engage in the debate around 'Responsible AI', a discussion around specific values is lacking, and the presence of utopian elements is limited to *control of nature*, *human safety* and *privacy*. Additionally, the dystopian element of *control of power* could be identified through interpretation taking into account the socio-cultural context and the role of corporations in dystopian visions involving technologies such as AI.

4.3 European Civil Law Rules in Robotics Study

This study was commissioned by the European Parliament's Legal Affairs Committee in order to provide a legal and ethical evaluation and analysis of possible future European civil law rules in robotics and AI. The study was carried out by the Policy Department C for Citizens' Rights and Constitutional Affairs, published on 12 October 2016.

> 'In this regard, it is essential that the big ethical principles which will come to govern robotics develop in perfect harmony with Europe's humanist values. The "Charter on Robotics", which was introduced in the draft report, moves in this direction.' (EU1, 2016, p. 7)

This excerpt clearly belongs into the morally responsible AI category, as by analyzing the text we are able to identify the sentence '(...) it is essential that the big ethical principles which will come to govern robotics (...)' (EU1, 2016, p. 7). In this case 'to govern robotics' can be interpreted as to ensure that the robot adheres to these ethical principles. In the same sentence the authors write 'in perfect harmony with

Europe's humanist values' (EU1, 2016, p. 7). Humanism being an ideology that values the human being, it is connected to utopian elements of human agency, more accurately human *liberty* and *justice*.

> 'In reality, advocates of the legal personality option have a fanciful vision of the robot, inspired by science-fiction novels and cinema. They view the robot - particularly if it is classified as smart and is humanoid - as a genuine thinking artificial creation, humanity's alter ego. We believe it would be inappropriate and out-of-place not only to recognise the existence of an electronic person but to even create any such legal personality. Doing so risks not only assigning rights and obligations to what is just a tool, but also tearing down the boundaries between man and machine, blurring the lines between the living and the inert, the human and the inhuman. Moreover, creating a new type of person - an electronic person - sends a strong signal which could not only reignite the fear of artificial beings but also call into question Europe's humanist foundations. Assigning person status to a nonliving, non-conscious entity would therefore be an error since, in the end, humankind would likely be demoted to the rank of a machine. Robots should serve humanity and should have no other role, except in the realms of science-fiction.' (EU1, 2016, pp. 15-16)

This lengthy excerpt provokes an intriguing discussion around the philosophical and societal implications of a legal personhood for robotics and had to be fully quoted in order to demonstrate and uphold the full content and context. Due to the coding rule *AI having the status of a legal personhood* this excerpt clearly belongs into the category of legally responsible AI, also because the discussion undeniable revolves around the topic. The authors state that those who call for a legal personality have 'a fanciful vision of the robot, inspired by science-fiction' (EU1, 2016, p. 15) including the notion that they would be 'genuine thinking artificial creation, humanity's alter ego.' (EU1, 2016, p. 16). In the following sentences the authors ridicule this fanciful vision by claiming AIs would be non-living entities devoid of consciousness. Given the fact that this document is supposed to represent a scientific study, this dogmatism without any kind of underlying scientific evidence is somewhat surprising, if not shocking. Scientific research is by no means so far that it would able to explain how consciousness is created (Gillies, 1996). Therefore, statements that negate the possibility of AI consciousness are factually wrong. We simply don't know what creates consciousness, therefore statements as to which entities are conscious or not are out of place and unscientific. The authors did not even cite any scientific discussions with regards to the possibility of artificial consciousness, when in fact there are plenty (Christian et al., 1997; Du & Li, 2007; Haugeland, 1997).

But what would be the implications if future AIs would indeed be thinking, conscious entities? Until now, this analysis has only focused on utopian and dystopian elements in relation to human utopias and dystopias. But wouldn't a future where artificial, consciousness entities have no basic rights be a dystopia that is characterized by human *domination*, AI *subjugation*, AI *oppression*, AI *discrimination* and basically AI *slavery*? Surely, these implications are fanciful and far-fetched. Nevertheless, they are based on one assumption only: The assumption that by creating artificial intelligence, we create artificial

consciousness. Given the fact that it is a theoretical possibility, one excerpt and paragraph was devoted to the discussion around dystopia. Not a human dystopia but a dystopia for conscious AI. While current AI systems have not reached the level of human intelligence yet, stronger AI (AGI, Superintelligence) - and therewith a possible higher probability of AI becoming conscious - may be closer than most of us would expect as some experts argue (Kurzweil 2005, 2010). Therefore, this discussion may not necessarily concern a remote, fictional future, but a tangible, real one. Yet, for this research no conclusions can be drawn due to the lack of scientific evidence around AI consciousness and corresponding implications. This is also due to the fact that science is concerned with objectivity while consciousness is inherently subjective. Nevertheless, this brief discussion is relevant in the context, since the entirety of our morals, values and ethics, including our idea of human responsibility, is developed, derived of and based on the premise of human consciousness.

4.4 Artificial intelligence - The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society

This document represents the own-initiative opinion by the European Economic and Social Committee, released on May 31, 2017.

> 'The EESC calls for a code of ethics for the development, application and use of AI so that throughout their entire operational process AI systems remain compatible with the principles of human dignity, integrity, freedom, privacy and cultural and gender diversity, as well as with fundamental human rights.' (EU4, 2017, p. 3)

In this case the category rule *ethical framework for the development/design of AI* applies. Here a code of ethics could be interpreted as an ethical framework, thus the statement belongs into the R1 category. In the same sentence, the author mentions the utopian elements of dignity, freedom and privacy which should be compatible with AI systems. The author further states that AI systems should be compatible with fundamental human rights. If the author refers to the Universal Declaration of Human Rights by the UN, it might be worthwhile to briefly analyze the declaration for utopian or dystopian elements. Using the pre-established category system indeed a number of utopian elements can be identified: *Equality* (Article 1), *Privacy* (Article 12) and *Safety* and *Liberty* (Article 3) (Assembly, 1948). Moreover, certain dystopian elements can be identified as well: Freedom from *Discrimination* (Article 2), Freedom from *Slavery* (Article 4) and Freedom from *Torture* (Article 5) (Assembly, 1948). Of course, one could argue that the formulation 'Freedom from…' means that humans should be protected from these things, however it

doesn't change the fact that these terms themselves represent dystopian elements, regardless of the context in which they appear.

> "The EESC calls for a European AI infrastructure consisting of open-source learning environments that respect privacy, real life test environments and high-quality data sets for developing and training AI systems. The EESC highlights the (competitive) advantage the EU can gain on the global market by developing and promoting 'responsible European AI systems', complete with European AI certification and labels." (EU4, 2017, p. 4)

This excerpt takes a special role in the context of this thesis since for the first time an institutional body of the EU calls for the development of 'responsible European AI systems' (EU4, 2017, p. 4). Thus, the statement can be allocated to the 'Responsible AI' category. In this context the committee calls for an AI infrastructure that respects the utopian element of *privacy*. Unfortunately, the committee does not further go into detail how such a responsible AI may look like and what features or characteristics would ensure its responsible. The committee does however give some hints without explicitly mentioning responsible AI systems, as the next excerpts will show.

> 'The AI systems now being developed will not have any built-in ethical values. We humans must make provision for them in AI systems and in the environments in which they are used. The development, application and use of AI systems (both public and commercial) must take place within the limits of our fundamental norms, values, freedoms and human rights. The EESC therefore calls for the development and establishment of a uniform global code of ethics for the development, application and use of AI.' (EU4, 2017, p. 6)

In the third sentence the committee mentions a 'global code of ethics for the development, application and use of AI' (EU4, 2017, p. 6). Once again, we may interpret this code of ethics as an ethical framework, thus the excerpt be allocated to the R1 category and belongs to the debate around the 'Responsible AI' concept. Within this statement the committee calls for human rights, freedoms, values and fundamental norms to determine the limits for the application and development of AI systems. Whereas *freedom* represents the code for the utopian element of *liberty*, the author does not further explain which fundamental norms are meant. Concerning the involvement of human rights, the same previously discussed utopian and dystopian elements as part of the UN declaration apply. The following excerpt stands in light of the headline 'Superintelligence' (EU4, 2017, p. 11)

> ‘The EESC calls for a human-in-command approach including the precondition that the development and application of AI be responsible and safe, where machines remain machines and people will be able to retain control over these machines at all times.’ (EU4, 2017, p. 11)

Due to the committee demanding the application and development of AI to be responsible, it can be allocated to the morally responsible AI category. Here we can identify the utopian elements of *safety* and *control over technology*. In the sentence before this excerpt, the committee states that “as a result, there are experts who opt for a ‘kill switch’ or reset-button, which we can use to deactivate or reset an out-of-control or superintelligent AI system” (EU4, 2017, p. 11). Here the dystopian element of *losing control over technology* can be identified. This clarifies why in the committee calls for humans to be in *control of technology* at all times, because according to various experts, an out-of-control superintelligence poses serious if not existential risks to humanity (Bostrom, 2002). In the light of such risks, the responsible application and development of AI becomes almost dramatically important.

4.5 Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems

This document represents an advisory report for the President of the European Commission by the European Group on Ethics in Science and New Technologies. It was published on March 9, 2018.

> ‘In recent debates about Lethal Autonomous Weapons Systems (LAWS) and Autonomous Vehicles there seems to exist a broad consensus that Meaningful Human Control is essential for moral responsibility. The principle of Meaningful Human Control (MHC) was first suggested for constraining the development and utilisation of future weapon systems. This means that humans - and not computers and their algorithms - should ultimately remain in control, and thus be morally responsible.’ (EU5, 2018, p. 9)

Applying the coding rules, we can assert that this discussion fits in the morally responsible AI category, due to the last sentence stating that humans and not algorithms should be morally responsible. The context of this discussion is the debate around Lethal Autonomous Weapon Systems (LAWS) which are basically weapons with integrated AI systems. In this debate several moral and ethical questions arise, for example should such LAWS be allowed to kill humans without a human giving the order? Or should such autonomous weapon systems exist at all, due to the serious security threats they pose in events where LAWS get hacked for example? The authors of this statement suggest that the principle of Meaningful Human Control - which was developed by a UK-based NGO called Article 36 (2016) - may be considered to constrain development and use of LAWS. Hence, this excerpt as well as the general debate around LAWS is very much concerned with the utopian element of *control over technology*, especially in relation

to the fundamental question of who should ultimately bear moral responsibility for the actions and inactions by such weapon systems.

From page 16 on the group discusses a number of ethical principles. Through interpretation we can assume that these principles stand in the context of development and application of AI systems. Thus, the following excerpts below belong to the debate around morally responsible AI.

> 'The principle of human dignity, understood as the recognition of the inherent human state of being worthy of respect, must not be violated by 'autonomous' technologies. This means, for instance, that there are limits to determinations and classifications concerning persons, made on the basis of algorithms and 'autonomous' systems, especially when those affected by them are not informed about them. It also implies that there have to be (legal) limits to the ways in which people can be led to believe that they are dealing with human beings while in fact they are dealing with algorithms and smart machines. A relational conception of human dignity which is characterised by our social relations, requires that we are aware of whether and when we are interacting with a machine or another human being, and that we reserve the right to vest certain tasks to the human or the machine.' (EU5, 2018, p. 11)

This rather long excerpt includes a high number of utopian elements, which party can only be identified to interpretation. The first utopian element of *dignity* can be easily identified by looking at the text. The second sentence discusses '(...) limits to (...) classifications concerning persons (...)' (EU5, 2018, p. 11). Here we may identify the utopian element of *impartiality*, since classification of people could lead entail discrimination. Thus, a limit to the classification of people could be demanded with the intention to ensure non-discrimination. Furthermore, the group advocates legal limits with regards to people being perceived to be talking to humans, when in fact they are communicating with an AI. Further interpreting the used term 'legal limits' could imply that in the future, a legally responsible AI may be legally obligated to notify a human that he or she is not communicating with another human, but with an AI. Looking at Google's latest *Duplex* technology (Leviathan & Matias, 2018), which is able to closely resemble human voices and communication behavior, this might soon become a reality in the near future. The last sentence of the excerpt contains the utopian element of *self-determination*, since the authors demand that humans should reserve the right to decide whether or not to 'vest certain tasks (...) to the machine' (EU5,2018, p. 11). As discussed earlier, the implication of not being able to allocate tasks to AIs after one's own will is that AIs will make the decisions for us, thereby reducing our ability to determine our own future. In the paragraph following the afore cited excerpt, the utopian elements of *liberty* and *control over technology* can be identified by looking at the text. Here the sentence "All 'autonomous' technologies must, hence, honour the human ability to choose whether, when and how to delegate decisions and actions to them." (EU5, 2018, p. 16) once again includes the utopian aspect of *self-determination*.

> "The principle of responsibility must be fundamental to AI research and application. 'Autonomous' systems should only be developed and used in ways that serve the global social and environmental good, as determined by outcomes of deliberative democratic processes. This implies that they should be designed so that their effects align with a plurality of fundamental human values and rights." (EU5, 2018, p. 16)

Here the utopian element of *environmental preservation* can be identified, since autonomous systems should serve environmental good which implies that the environment should not be damaged. Hence, after the opinion of the group a morally responsible AI is also an AI that protects the environment. Aside from the aforementioned elements the group also mentions the utopian elements of *justice*, *safety* and *privacy* (EU5, 2018, pp. 18-19) as part of the discussion around ethical principles for the development and design of AI.

4.6 Artificial Intelligence for Europe

The final document to be analyzed is the communicated from the European Commission to the European Parliament and the European Council, the Council, the European Economic and Social Committee as well as to the Committee of the Regions with a total length of 20 pages. Because it was published on April 25, 2018 it represents the most recent EU document on the debate around responsible AI. In this document the Commission announced that the "guiding principle of all support for AI-related research will be the development of 'responsible AI', putting the human at the centre (...)" (EU6, 2018, p. 8)

> 'As with any transformative technology, some AI applications may raise new ethical and legal questions, for example related to liability or potentially biased decision-making. The EU must therefore ensure that AI is developed and applied in an appropriate framework which promotes innovation and respects the Union's values and fundamental rights as well as ethical principles such as accountability and transparency.' (EU6, 2018, p. 3)

While the syntax '(...) some AI applications may raise new ethical and legal questions (...)' (EU6, 2018, p. 3) is somewhat vague, the subsequent examples of biased decision-making and liability clarify that the discussion revolves around 'Responsible AI' and that the excerpt can be allocated to both categories, morally as well as legally responsible AI. In this case the keyword *ethical* alone would not have been sufficient to determine without a doubt that the discussion concerns the *Ethical AI* code, and therewith category R1, however, the example of biased decision-making clarifies that the sentence indeed discusses actions by an AI and not by some other actor. Moreover, the following sentence confirms this conclusion

due to the wording '(...) that AI is developed and applied in an appropriate framework which (...) respects (...) ethical principles such as accountability (...)' (EU6,2018, p. 3). As defined in the coding rules, text discussing an *ethical framework for the application of AI* belongs into the morally responsible AI category. Here a slight discrepancy between the established theoretical and categorical framework can be identified. During the categorization of this research the term *accountability* was attributed to the legally responsible AI category, because the meaning of the word in a legal context is closely related to that of *liability* (Accountable, 2016). However, another synonym of being accountable is being responsible for one's actions (Accountable, 2016) which explains why the European Commission uses accountability as an example of an ethical principle. Within this excerpt, two instances of text include utopian elements. First, the phrase 'The EU must therefore ensure that AI is developed and applied in an appropriate framework which promotes innovation (...)' (EU6, 2018, p. 3) can be attributed to the utopian element of *progress* due to the emphasis on innovation. Innovation as such has two meanings: The introduction of something new as well as a new idea, method, or device (Innovation, 2016). In relation to AI innovation thus means technological progress, which is central to techno-utopian ideology where advances in technology bring salvation to mankind (Dinello, 2005). The second instance where utopian elements can be identified follows the question what ethical principles should be included in this ethical framework. Here the Commission refers to Article 2 of the Treaty on the European Union:

> 'The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail.' (Treaty of Lisbon, 2007).

Here a number of utopian elements can be identified, manifested in the law of the EU: Human *dignity*, *liberty* (*freedom*), *impartiality*, *justice* and *equality*. Additionally, it states that human rights should be respected. As the analysis of the opinion by the European Economic and Social Committee has shown, human rights themselves include a number of utopian elements along with a number of dystopian elements, whereas the dystopian elements stand in the light of prohibition.

> '(...) Europe should strive to increase the number of people trained in AI and encourage diversity. More women and people of diverse backgrounds, including people with disabilities, need to be involved in the development of AI, starting from inclusive AI education and training, in order to ensure that AI is non-discriminatory and inclusive.' (EU6, 2018, p. 13)

This excerpt includes aspects of morally responsible AI and the utopian element of *impartiality* due to the phrasing 'in order to ensure that AI is non-discriminatory' (EU6, 2018, p. 13). In this case the Commission calls for the development of AI to adhere to an ethical aspect, namely not to discriminate against people of different genders, disabilities or ethnicities. AI systems act depending on which data was used for their development. If the input data was incomplete or was derived from potentially biased human sources, this can lead to the AI output being biased and discriminatory. Therefore, an important aspect of making an AI responsible means that the data which is used to train and design AI has to be objective and comprehensive. Due to this reason, the Commission calls for more women and diversity in the (responsible) development of AI, in order to ensure that AI systems are impartial and thus act ethically responsible (EU6, 2018).

The excerpt below has the overarching headline 'Ensuring an appropriate ethical and legal framework' (EU6, 2018, p. 14). Moreover, the first sentence below the headline reads 'An environment of trust and accountability around the development and use of AI is needed' (EU6, 2018, p. 14). Judging from this, we can assume that the ethical and legal framework concern the development and design of AI. Therefore, the excerpt below stands in context with the development of 'Responsible AI' and can be categorized as such according to the corresponding coding rules *ethical framework for the development/design of AI* and *legal framework for the development/use of AI*. Moreover, the excerpt itself contains the code *AI ethics*:

> 'As a first step to address ethical concerns, draft AI ethics guidelines will be developed by the end of the year, with due regard to the Charter of Fundamental Rights of the European Union. The Commission will bring together all relevant stakeholders in order to help develop these draft guidelines. The draft guidelines will address issues such as the future of work, fairness, safety, security, social inclusion and algorithmic transparency. More broadly, they will look at the impact on fundamental rights, including privacy, dignity, consumer protection and non-discrimination. They will build on the work of the European Group on Ethics in Science and New Technologies and take inspiration from other similar efforts.' (EU6, 2018, pp. 15-16).

Since the excerpt concerns the debate around 'Responsible AI' we may now look for utopian or dystopian elements. Through textual analysis we can identify a number of utopian elements here. First of all, the draft guidelines according to which AI shall be developed and used will by declaration of the Commission address the utopian elements of *safety*, *privacy*, *dignity* and *impartiality* (non-discrimination). Furthermore, the Commission states that these ethical guidelines shall be developed with regards to the Charter of Fundamental Rights of the EU. Taking a look at the content of this charter we can identify a number of additional utopian elements. Aside from the aforementioned elements, the charter entails additional utopian elements of *liberty*, *equality* and *justice* (European Union, 2010). What is more, the

first title of the charter (Dignity) also includes the prohibition of the dystopian elements of *torture* and *slavery* (European Union, 2010).

The last sentence of the afore cited excerpt 'They [the draft guidelines] will (...) take inspiration from other similar efforts.' (EU6, 2018, p. 16) refers to the paper by the European Group on Ethics in Science and New Technologies, as well as to three examples of international efforts via a footnote: The Asilomar AI principles (Future of Life Institute, 2017), the Montréal Declaration for Responsible AI draft principles (University of Montréal, 2017) and the UNI Global Union Top 10 Principles for Ethical AI (UNI Global Union, n.d.).While the paper by the European Group on Ethics in Science and New Technologies has already been analyzed earlier in this paper, the three international projects could add some additional depth and elements to this analysis. After all, the Commission has clearly stated that it will take inspiration from these efforts, which means that they might will play a role in future legislation. Whether this role will be significant or not remains to be seen. But do these international efforts add additional utopian or dystopian elements to the aforementioned ones? In most cases they resemble the same utopian elements discussed in the paragraph before, thus recurring elements will not be mentioned. However, principle number 16 and 22 of the Asilomar AI principles add some new elements. Point 16 states that 'Humans should choose how and whether to delegate decisions to AI systems (...)' (Future of Life Institute, 2017). The notion that humans and not AI systems should be the actor that decides whether and how AI systems should make decisions evokes the utopian element of *self-determination*. If we would not be able to determine what an AI should decide for us and what not, the extent to which we would be able to determine our own future would be severely restricted. This issue may become difficult regarding the use of AI by people with disabilities. On the one hand, AI will be able to expand possibilities of self-determination by complementing what peoples with disabilities are missing. On the other hand, it will be a challenge to design a responsible AI in such a way that the lives of people with mental disabilities for example won't be dictated by algorithms.

Principle 22 states that 'AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.' (Future of Life Institute, 2017). As discussed in the theory section, recursive self-improvement is the process by which an AI could (theoretically) exponentially increase its capability and thus become a superintelligence (Kurzweil, 2005; Yampolskiy, 2015). Therefore, the utopian elements of *safety* and *control over technology* here are intended to guarantee that it will not damage or threaten humanity and that ultimately, humanity is in control of such a powerful AI. Whether it is actually possible to ensure safety, value alignment and to uphold absolute control over such powerful technology is in the eyes of many experts in the field not certain (Bostrom, 2014; Tegmark, 2017). While the UNI Global Union principles add no utopian or dystopian elements to the discussion by the Commission, the Montréal

Declaration for Responsible AI adds the utopian element of *liberty* due to the use of the term *autonomy* as one of their main principles for responsible AI (University of Montréal, 2017).

5. Conclusion

The qualitative content analysis of the six EU documents after Mayring (2014) has revealed several elements of 'Responsible AI' concept as well as of utopian and dystopian visions. Concerning the first sub-question of this research 'Which particular elements of 'Responsible AI' are represented in the data?' the analysis has shown that both theoretical elements of the concept, legally as well as morally responsible AI were significantly represented in the debate throughout all documents. In most cases they could be identified through coding rules that included legal, ethical and value guidelines or frameworks for the development and design of AI. Only in very few instances, the authors directly used the term 'Responsible AI' such as in the documents by the EESC or the Commission. However, since the European Commission has recently declared 'Responsible AI' to be the guiding principle 'of all support for AI-related research' (EU6, 2018, p. 8) it is possible that in the future, more EU actors will adopt the term in discussions around AI instead of transcribing it using the aforementioned elements. Concerning the second subquestion "How are elements of 'Responsible AI' linked to elements of utopian and dystopian visions?" the answer is manifold. A strong connection was identified between legally responsible AI and the utopian element of *control over technology*. This is due to the fact that if AI systems cause damage, the question of who is legally liable for the damage is very much concerned with whether the AI system is under control of a human and who then becomes the legally responsible actor. Another significant link was found between morally responsible AI and the utopian elements of *safety*, *dignity*, *impartiality* and *self-determination*. Throughout the analysis these connections occurred repeatedly and can be attributed to the fact that our idea of moral responsibility is strongly connected to our fundamental values and human rights. There are many ways through which AI systems may threaten our values and rights through actions or inactions, hence a 'Responsible AI' could ensure that values and rights are respected and that an AI behaves accordingly. On the other hand, the many ways through which AIs may act irresponsibility, by acting discriminatory for instance, provoke the requirement for a responsible AI in the first place.

Concerning the answer to the main research question "Which elements of utopian and dystopian visions are present in the debate about 'Responsible AI' in the European Union between 2016 and 2018?" many different elements could be identified. The utopian elements present are *dignity*, *liberty*, *safety*,

privacy, *justice*, *impartiality*, *equality*, *progress*, *self-determination*, *control of technology*, *control of nature* and *environmental preservation*. The identified dystopian elements are *war*, *inequality*, *discrimination*, *mass-surveillance*, *control of power* and *losing control over technology*. Here it is important to emphasize ones again that the process by which the categories of the 'Responsible AI' concept as well as the utopian and dystopian visions were developed followed theoretical findings in a deductive category formation process after the methods by Philipp Mayring (2014). Hence, different theoretical foundations as well as a different conceptualization of the 'Responsible AI' concept may lead to different results. Nevertheless, since the establishment of the categories followed a deductive process, the degree to which the results are intersubjective and degrees of construct validity and reproducible are high. But what exactly does this research contribute to the scientific, political and societal realm? As for the scientific world, something like a 'Responsible AI' has never been conceptualized before. Therefore, this conceptualization stands in the light of explorative research in relation to concepts involving artificial intelligence. Moreover, a lot of literature exists on the topic of AI in dystopia and utopian visions. Yet, a theoretical connection between responsible AI and utopian and dystopian visions along with comprehensive list containing elements of utopian and dystopian elements in connection to AI has not been developed before. But this research does not only add theoretical an understanding of the connection between AI, responsibility as well as utopian and dystopian visions. The answer to the main research question adds tangible knowledge regarding the presence of utopian and dystopian elements in the debate about AI in the EU. The results show that a high number of utopian elements is present in this debate, while some dystopian elements could be identified as well. This might be a sign that making AI responsible is not only contributing to the attainment of utopist ideals but may be a necessity. The potential of AI - especially considering the rapid, almost exponential technological progress in AI research - to radically transform our societies is undeniable. The ultimate question is, will it transform society for better or worse? And how can we ensure as a society to promote the positive effects of AI while minimizing risks?

The analysis in this research has shown that the EU is very much concerned with this topic. The high number of utopian elements in the debate around 'Responsible AI' show that the EU's strategy to make AI responsible is highly characterized by utopian aspirations, while the presence of dystopian elements show that AI could also be big threat to society, especially regarding stronger, future versions of AIs such as a theoretical AGI or a Superintelligence and that EU actors are aware of this which is a good sign. But of course, this research is limited in that it focused on the debate around the 'Responsible AI' concept. Due to this focus, those utopian and dystopian elements which have weak or no links to AI responsibility were left out of the discussion. Perhaps one of the debates where utopian and dystopian elements would play a significant role in relation to AI is the debate around the socio-economic impact of

AI on contemporary economic inequalities. Thus, further research could analyze this debate without limiting the focus on responsible AI. Furthermore, future research could analyze the political debates around ‘Responsible AI’ in other countries in order to compare results and draw conclusions. This would be especially interesting with cases like the US or China. While we in Europe may take certain utopian ideals for granted, due to their presence in the EU charter of fundamental rights for instance, research focusing on other countries and cultures could potentially lead to much different results in terms of the visibility of utopian and dystopian elements. While the EU holds human rights such as privacy in very high regards, the situation in China for example is different. Hence, further research involving the identification of utopian and dystopian elements in the debate about ‘Responsible AI’ may enable us to make more meaningful conclusions, since research in different regions of the world, with different languages, morals, values and laws may lead to very different results, possibly involving a very different understanding of what responsibility is, and what a ‘Responsible AI’ would look like in the eyes of different cultures. The European Commission stated that ‘The way we approach AI will define the world we live in.’ (EU6, 2018, p. 2). The results of this thesis suggest that this approach is significantly characterized by utopian elements, indicating that the development of a ‘Responsible AI’ may contribute to a more utopian future.

References

Accountable. (2016). In *Merriam-Webster's dictionary* (13th ed.). Springfield, MA: Merriam-Webster.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251-261.

Arkin, R. C. (2010). The case for ethical autonomy in unmanned systems. *Journal of Military Ethics*, *9*(4), 332-341.

Article 36. (2016). *Key elements of meaningful human control*. Geneva, Switzerland.

Assembly, U. G. (1948). Universal declaration of human rights. *UN General Assembly*.

Ashrafian, H. (2015). Artificial intelligence and robot responsibilities: Innovating beyond rights. *Science and engineering ethics*, *21*(2), 317-326.

Bostrom, N. (2002). Existential Risks. *Journal of Evolution and Technology*, 9(1), 1-31.

Bostrom, N. (2014). The Control Problem. *Superintelligence*. Oxford University Press 127-144.

Brian A. Garner, editor in chief. (2014). *Black's law dictionary.* St. Paul, MN: Thomson Reuters.

Bryson, J. J., Caliskan, A., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data: complementary research strategies*. Sage Publications, Inc.

Christian, W., Franklin, S., McKay, S. R., & Wolpert, S. (1997). Artificial minds. *Computers in Physics*, *11*(3), 258-259.

Dinello, D. (2005). *Technophobia!: science fiction visions of posthuman technology.* University of Texas Press.

Du, Y., & Li, D. (2007). *Artificial intelligence with uncertainty*. Chapman and Hall/CRC.

European Commission. (2018). *Artificial Intelligence for Europe*. Office for Official Publications of the European Communities. Retrieved from: https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

European Economic and Social Committee. (2017). *Artificial intelligence - The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society*. Office for Official Publications of the European Communities. Retrieved from: https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/artificial-intelligence

European Parliament. (2017). *Report with recommendations to the Commission on Civil Law Rules on Robotics*. Office for Official Publications of the European Communities. Retrieved from http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A8-2017-0005+0+DOC+XML+V0//EN

European Parliament. (2016). *European Civil Law Rules in Robotics Study*. Office for Official Publications of the European Communities. Retrieved from: http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf

European Parliament. (2016). *Artificial Intelligence: Potential Benefits and Ethical Considerations*. Office for Official Publications of the European Communities. Retrieved from: http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/571380/IPOL_BRI%282016%29571380_EN.pdf

European Union. (2010). *Consolidated Versions of the Treaty on European Union and of the Treaty on the Functioning of the European Union: Charter of Fundamental Rights of the European Union*. Office for Official Publications of the European Communities.

Eshleman, E. (2014). Moral Responsibility. *Stanford Encyclopedia of Philosophy*. Retrieved March 12, 2018 from: https://plato.stanford.edu/entries/moral-responsibility/

Friedman, B., & Kahn Jr, P. H. (1992). Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software*, *17*(1), 7-14.

Future of Life Institute. (2017). Asilomar AI principles. Retrieved from: https://futureoflife.org/ai-principles/?cn-reloaded=1

Gillies, D. (1996). Artificial intelligence and scientific method.

Goertzel, B. (2013). Ben Goertzel on AGI as a Field. *Machine Intelligence Research Institute*. Retrieved March 12, 2018 from: https://intelligence.org/2013/10/18/ben-goertzel/

Haugeland, J. (Ed.). (1997). *Mind design II: philosophy, psychology, artificial intelligence*. MIT press.

Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and information technology*, 16(3), 197-206.

Innovation. (2016). In *Merriam-Webster's dictionary* (13th ed.). Springfield, MA: Merriam-Webster.

Kohlbacher, F. (2006, January). The use of qualitative content analysis in case study research. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* (Vol. 7, No. 1).

Kurzweil, R. (2005). Superintelligence and Singularity. *The singularity is near: When humans transcend Biology*, 7-33. Viking.

Kurzweil, R. (2010). *The singularity is near*. Gerald Duckworth & Co.

Leviathan, Y., Matias, Y. (2018). Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone [Blog post]. Retrieved from: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

Lin, P., Bekey, G. A., & Abney, K. (2009). Robots in war: issues of risk and ethics. In: R. Capurro & M. Nagenborg, *Ethics and Robotics*. Heidelberg: AKA Verlag, 49-67.

Nilsson, N. J. (2007). The physical symbol system hypothesis: status and prospects. In *50 years of artificial intelligence*, 9-17. Springer, Berlin, Heidelberg.

Mayring, P. (2014). Qualitative content analysis: theoretical foundation, basic procedures and software solution

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, *6*(3), 175-183.

Miłkowski, M. (2013). Reverse-engineering in Cognitive-Science. In Marcin Miłkowski & Konrad Talmont-Kaminski (eds.), *Regarding Mind, Naturally*. Cambridge Scholars Press. pp. 12-29.

Schneider, S. (2016). *Science fiction and philosophy: from time travel to superintelligence*. John Wiley & Sons.

Swierstra, T. (2015). Introduction to the Ethics of New and Emerging Science and Technology. *Handbook of Digital Games and Entertainment Technologies*, pp.1-25, doi:10.1007/978-981-4560-52-8_33-1.

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf.

The European Group on Ethics in Science and New Technologies. (2018). *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*. Office for Official Publications of the European Communities. Retrieved from: http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf

Treaty of Lisbon. (2007). *European Union.* Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:12007L/TXT

UNI Global Union. (n.d.). 10 Principles for Ethical AI. Retrieved from: http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/

University of Montréal. (2017). The Declaration. Retrieved from: https://www.montrealdeclaration-responsibleai.com/the-declaration

Winner, L. (1997). Technology today: Utopia or dystopia?. *Social research*, 989-1017.

Wolcott, H. F. (1994). *Transforming qualitative data: Description, analysis, and interpretation*. Sage.

Yampolskiy, R. V. (2015). From seed AI to technological singularity via recursively self-improving software. *arXiv preprint arXiv:1502.06512*.

Appendix A: Category system

Category R: Responsible AI

Category definition: Category definition: The categorization of the 'Responsible AI' concept follows the insights from the theory section. Corresponding with the conceptualization, 'Responsible AI' is divided into two subcategories: Morally responsible AI as well as legally responsible AI. Both subcategories have an or relationship, meaning that the identification of one of the two dimensions is sufficient for the 'Responsible AI' concept to enter into force.

Dimension	Anchor example	Coding rule
Subcategory R1: Morally responsible AI	'whereas the Union could play an essential role in establishing basic **ethical principles** to be respected **in the development, programming** and use **of robots and AI** and in the incorporation of such principles into Union regulations and codes of conduct, with the aim of shaping the technological revolution so that it serves humanity and so that the benefits of advanced robotics and AI are broadly shared (...)' (EU3, 2017, p. 6)	Responsible AI (moral/ethical/value context) **or** AI responsibility (moral/ethical/value context) **or** Adherence of AI to social responsibility **or** Socially responsible AI **or** Adherence of AI to ethics **or** Ethical AI **or** AI ethics **or** Adherence of AI to values **or** Ethical/value principles/framework for the development/design of AI
Subcategory R2: Legally responsible AI	'Believes that robotics and **artificial intelligence**, especially those with built-in autonomy, including the capability to independently extract, collect	AI being subject to laws **or** AI legislation

	and share sensitive information with various stakeholders, and the possibility of self-learning or even evolving to selfmodify, **should be subject to** robust conceptual **laws** or principles, such as that a robot may not kill or harm a human being and that it must obey and be controlled by a human being;' (EU3, 2017, p. 36).	**or** AI being legally responsible **or** Responsible AI (legal context) **or** AI having the status of a legal personhood **or** Electronic personhood **or** AI liability/accountability/obligation **or** Legal framework for the development/design/use of AI

Category V: Elements of utopian and dystopian visions

Category definition: This category includes the various elements of utopian and dystopian visions that were discussed in the theory section. It consists of two subcategories: Elements of utopian visions and elements of dystopian visions. The elements which can be identified through the coding rules all have an *or* relationship meaning that the presence of one of the elements is sufficient for the corresponding category to apply.

Dimension	Anchor example	Coding rule
Subcategory V1: Elements of utopian visions	'The EESC calls for a code of ethics for the development, application and use of AI so that throughout their entire operational process AI systems remain compatible with the principles of human **dignity**, integrity, **freedom**, **privacy** and cultural and gender diversity, as well as with fundamental human rights.' (EU4, 2017, p. 3)	Control of Nature **or** Control of Technology **or** Revitalization of politics **or** Dignity **or** Privacy **or** Impartiality (non-discrimination) **or** Equality **or** Liberty (Freedom/Autonomy) **or** Self-determination **or** Personal fulfillment **or** Progress **or** Peace **or**

		Harmony **or** Life enhancement **or** Individual and societal redemption **or** Environmental preservation
Subcategory V2: Elements of dystopian visions	'Highlights the importance of preventing **mass-surveillance** through robotics and artificial intelligence technologies;' (EU3, 2017, p. 37)	War **or** Chaos **or** Poverty **or** Control of power **or** Losing control over technology **or** Oppression/Suppression **or** Mass-surveillance **or** Social/economic inequalities **or** Domination **or** Discrimination **or** Subjugation **or** Slavery **or** Environmental destruction **or** Abolishment of rights

Appendix B: Corpus

Name and date of document	Code	Authorizing institution	Status	Length
European Civil Law Rules in Robotics Study (October 12, 2016)	EU1	European Parliament; Policy Department C for "Citizens' Rights and Constitutional Affairs"	Study	34 Pages
Artificial Intelligence: Potential Benefits and Ethical	EU2	European Parliament; Policy Department C for	Request to the Commission for submission of proposals	8 Pages

Considerations (October 12, 2016)		"Citizens' Rights and Constitutional Affairs"	according to Rule 46 of the Rules of Procedure	
Report with recommendations to the Commission on Civil Law Rules on Robotics (January 27, 2017)	EU3	European Parliament	Own-initiative opinion according to Rule 29(2) of the Rules of Procedure	64 Pages
Artificial intelligence - The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society (May 31, 2017)	EU4	European Economic and Social Committee	Advisory report to the President of the European Commission	13 Pages
Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems (March 9, 2018)	EU5	The European Group on Ethics in Science and New Technologies	Advisory report to the President of the European Commission	24 Pages
Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems (March 9, 2018)	EU6	European Commission	Communication	20 Pages

Appendix C: General content-analytical procedural mode by Mayring (2014)