Where is the “Ethos” in Artificial Intelligence codes of conduct?

Georgios Natsios
20191924

G. Natsios

Aalborg University Copenhagen
A.C. Meyers Vænge 15
2450 Copenhagen SV

AALBORG UNIVERSITY
STUDENT REPORT

P10 Thesis Project in
MSc. Techno-Anthropology

Standard Page Numbers
53,5

Character count
128.423

Date of Completion November 17, 2020

Title
"Where is the "Ethos" in Artificial Intelligence codes of conduct."

Project Period
February-November 2021

Participants
Georgios Natsios (20191924)

Supervisor
Tom Børsen

Copies: 1

Abstract

In this thesis report, I seek to analyze the codes of conduct that recommend principles for the ethical development of Artificial Intelligence, leading to a twofold target. Firstly, understanding the content of the codes of conduct and correspondingly identifying the ethical principles of AI and secondly, analyze these principles by techno-anthropological methods (combination of theoretical framework and digital methods) to decipher their meaning and outline recommendations on the development of the AI Ethical framework. To outline my recommendations, I am underlining Aristotelian and Foucauldian ethical notions, to identify the practices (ethos), that will assist on the sufficient exercise of ethical principles in the design and development of Artificial Intelligence systems.

Contents

1. Introduction

Globally, the problematization of the utilities and the unrestricted development of Artificial Intelligence is becoming more intense. Bias in algorithms, misuses of the technology, disinformation, exclusion of certain groups are a few of the examples that Artificial Intelligence has provoked the last few years (Royer, 2020). However, Artificial Intelligence is a technological development that may create several benefits for humanity and the environment, in various sectors such as medicine, the future of work, and sustainable development, among others (EPRS, 2020). Nevertheless, there should be cautions in the process.

In the latest decade, ethical development and design of Artificial Intelligence have been included in the dialogue and the discussions as an essential part of the development of this technology. As a result, a high number of research papers or codes of conduct, presenting potential AI Ethical principles have been published in the last ten years (Jobin et al, 2019). However, these codes of conduct seem to recycle specific theories and principles, avoiding practicalities or technical codes and precautions. Additionally, they seem to hunt a universal AI Ethical framework, omitting the cultural and social aspects of humanity.

Going through this research study, you will go across the "Ethos" "quest" of Artificial Intelligence codes of conduct. The definition of the word "Ethos" is going back to Ancient Greece and Aristotelian ethics, referring to a habit, a practice of an ethical action which can lead to the acquirement of ethical virtues (Athanasopoulos, 2013-2014). In this research, the word "Ethos" is a reference to "Virtue Ethics", which are overlooked from the contemporary conversations about ethical principles of Artificial Intelligence and a reference to the practical application of ethics that is lacking on the AI framework as I will show in the next chapters of this project.

By navigating through this research, the reader will first understand the different interpretations of Artificial intelligence and the problematization of the research project. Subsequently, to respond to the research questions, I will utilize a unique methodology coming from the techno-anthropological studies, combining a digital methods environment and classical ethical theories, with a focus on Aristotelian and Foucauldian ethics. Finally, you will observe the results and the suggestions of this research project in the analysis and discussion parts.

1.1. Artificial Intelligence

What is Artificial Intelligence or what kind of technology it is? Definitions are evolving constantly, but we must consider that it is a field that is not stationary but distributed to separate sectors and disciplines, it is movable all the time. Going back in time, Alan Turing was mentioning that a machine should be considered "intelligent" whenever a human who has a connection with it, cannot recognize if it is a human or a machine (Royer, 2020). Since AI is being distributed to different sectors, there are plenty of descriptions of this broad technology. However, in this research, I will outline some of the definitions and practices of AI, which will make it easier for the reader to reflect on the methods and utilities of this technological revolution.

Gasser and Almeida for instance, establish that, one cause for the difficulty of defining AI from a technical perspective, is that AI is not a single technology, but rather "a set of techniques and subdisciplines ranging from areas such as speech recognition and computer vision to attention and memory, to name just a few." (Gasser & Almeida as cited in Larsson 2020:439). While the majority assume that AI is a completely new technology, innovated in the 21st century, this impression is quite misguided. AI has existed many years ago, being around in the 1950s after the second world war, with the first conference in 1956 held at Dartmouth College, whereas researchers unveiled the term "Artificial Intelligence" and in an extremely positive view, they were implying that "Artificial Intelligence" would be developed in less than a generation, with the specific example of AI innovator Marvin Minsky announcing that "In from three to eight years we will have a machine with the general intelligence of an average human being" (CSSF, 2018:8). Ultimately, the development and progress of "Artificial Intelligence" were quite slower than expected.

Another inadequacy to define AI is the fact that even human intelligence is an unclear term. For instance, the Financial Stability Board has defined Artificial intelligence as "The theory and development of computer systems able to perform tasks that traditionally have required human intelligence." (CSSF, 2018:6). Thus, AI is having the role of executing intelligent tasks, such as the following,

- Problem solving and Reasoning
- Perception
- Learning
- Planning
- Ability to understand language and speech
- Ability to manipulate and move objects

At the same moment, AI can practice more than solely one task, delivering a multitasking result or just supporting humans in the decision-making process, whereas there are the ones who can take their own decisions to accomplish their tasks that are called "autonomous systems".

Another definition that is quite popular for AI is the one that stated from the EU Commission's communication on AI in Europe, in April 2018:

> "Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions—with some degree of autonomy—to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications). (The High-Level Expert Group, 2019b, p. 1, as cited in Larsson, 2020:440)"

It is interesting that this definition is referring to the autonomy -some degree of autonomy- of the AI systems. I will explore the principles of autonomy on a broader spectrum later in this thesis. Nevertheless, the important matter here is that we can witness AI software systems as well as physical machines, robots (software and hardware devices) and this is the extensive variety of AI applications that an AI ethical framework should be implemented.

With this mentioned and with the knowledge of the various applications of AI and the division with Machine Learning, Larsson is developing the aforementioned definition here:

> "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions." (Larsson, 2020:441)

As Larsson describes "There are thus differing aspects of AI to be considered in the definition of AI, as a challenge to the regulation, where the most central ones for today's development and use of AI tend to concern (1) autonomy/agency, (2) self-learning from large data amounts (or "adaptability"), and (3) the degree of generalizable learning." (Larsson, 2020:441). Hence, an AI software/system definition may be arguably distinctive due to its specific utility and capabilities.

Nevertheless, it is becoming evident from this chapter that a general definition of AI is undeniably complex, especially when the term intelligence is emerging into the discussion. However, the different areas and sectors of "Artificial Intelligence" can offer us the opportunity to understand the role and the practices of this technology that nowadays is being theorized as one of the cores of economic and political development. In the following visualization, the reader can explore the different subfields of AI, before starting the discussion about the ethical principles that should prevent the unrestrained development of these subsectors.

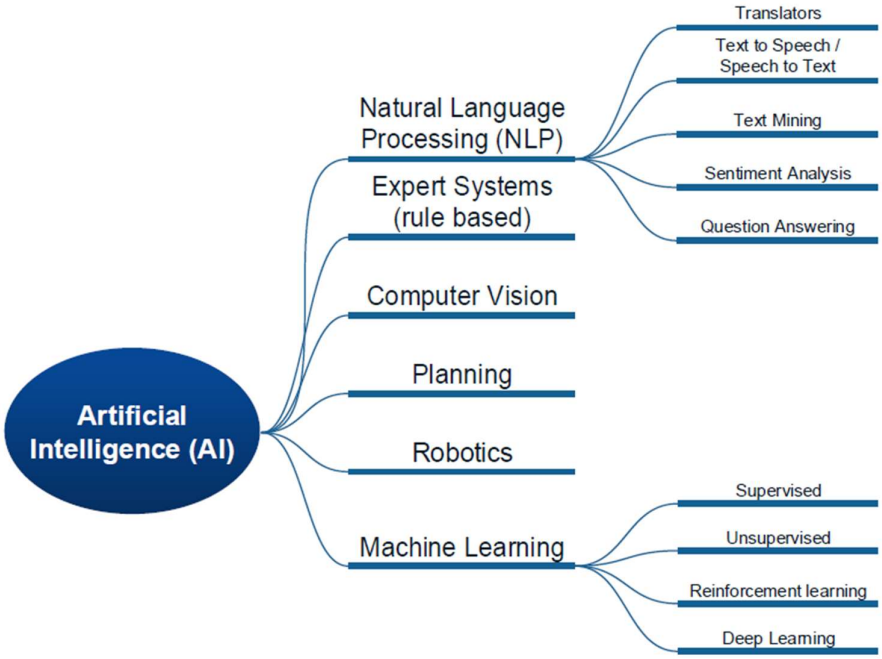


Figure 1, Source: CSSF, 2020:11. Main AI Subfields.

1.2. Artificial Intelligence utilities

Depending on the previously mentioned definitions, Artificial Intelligence seems like a technology with infinite capabilities, that can assist humans in different sectors from, daily tasks to healthcare, economy, business sectors, ecology and more. Before we dig into the importance of restricting the development of this technology within a responsible and ethical framework, I should refer to a few cases, that can prove how beneficial this technology may be under the right circumstances. The next cases will focus on Artificial Intelligence examples that are reflected as beneficial or helpful for humanity.

I will start with The World Bee Project[1], which is a project created for saving and protecting pollinators, people and the planet. An unknown fact for this project is that it uses artificial intelligence and more specifically,

"The World Bee Project hopes to learn how to help bees survive and thrive by gathering data through internet-of-things sensors, microphones, and cameras on hives. The data is then uploaded to the cloud and analyzed by artificial intelligence to identify patterns or trends that could direct early interventions to help bees survive." (Marr, 2020)

In the same direction, another Artificial Intelligence application is developed by Huawei, to help blind people, "see" emotions by translating sounds. The app, Face emotions

[1] https://worldbeeproject.org/. "The World Bee Project". n.d. Accessed November 4, 2021.

[2]"…translates these seven universal human emotions – anger, fear, disgust, happiness, sadness, surprise and contempt – into seven distinct sounds. The detected emotion is played through the phone's microphone." (Pownall, 2019). Generally, these two applications of Artificial Intelligence seem beneficial to certain groups, assisting their survival and making their life easier, proceeding to the responsible development of AI technology. Furthermore, attempts of confronting the global warm crisis, helping education, finance and healthcare are being constantly developed under Artificial Intelligence software (ITU, 2021).[3]

However, I will continue with some of the recent most popular forms of Artificial Intelligence, that is coming in the form of digital personal assistants, such as Amazon's Alexa, Apple's Siri and Microsoft's Cortana. These digital personal assistants can be integrated into a smartphone, laptop or have their form as a speaker, helping a person in various ways. For instance, they may assist a person in daily tasks, such as phone calls, alarms, reminders, music, games, etc. by asking permission on personal data and preferences. One can argue that these digital personal assistants may indeed assist with various daily tasks, providing a comfortable environment for the user.

Nevertheless, there are controversies on how personal data and privacy are being utilized by these companies. In the article of the Washington Post by Amy B. Wang, it is being described a situation during a murder, when then Amazon Echo, could record everything as a part of its recording AI system that is being kept in the data storages of Amazon and not being "utilized" as Amazon argues. Thus, in this case, Amazon denied providing the stored data recording, since it would have been a violation of the privacy and personal data of the user (potential murderer in this case) (Wang, 2017). I will return in this discussion regarding autonomy, privacy, transparency, responsibility and personal data, later in this project, under the analysis and discussion parts of this project, as well as more biased AI cases, concerning fairness, equality and security.

Nevertheless, the violation, in this case, is that the Amazon Alexa AI system actually can violate a person's privacy by recording sounds, however, not share these recordings under legal obligations. Hence, this Artificial Intelligence system, should have functioned under a specific AI Ethical framework, respecting specific guidelines and principles, regarding person's privacy in this case. Therefore, we can see that even though there are discussions and attempts for a responsible development of AI, still there are practical issues of implementation of ethics in Artificial Intelligence systems, that arises questions in which I will discuss in the next chapter of my problem analysis.

[2] https://www.dezeen.com/2019/01/02/huawei-app-blind-facing-emotions/. "Face emotions". January 2019. Accessed, November 4, 2021.
[3] https://www.itu.int/en/mediacentre/backgrounders/Pages/artificial-intelligence-for-good.aspx. "Artificial Intelligence for good". June 2021. Accessed, November 4, 2021.

2. Problem Analysis

In this chapter, I am going to outline the analysis of my overall problematization and how I concluded to the problem statement of this thesis project, starting with my literature search.

2.1. Literature search

This thesis project is using the Aalborg University Library's model of the 5 W's "what, where, words, work, wow" (See figure 2), as an inspiration of drawing on "Success, a Structured Search Strategy: Rationale, Principles, and Implications" (Zins, 2000).

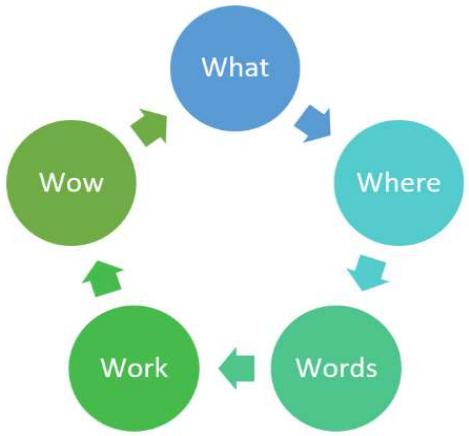


Figure 2, "The 5 W's"

The literature search took place as follows. Starting with the "what", the overall problem of the thesis project: "What are the challenges of the Ethical AI?". Moving on, to the "where", the academic database that the research took place was "Ebscohost", chosen for its ability to include a variety of different academic databases, which is given a wide option of publications. Subsequently, the "words", in which I used the following equation: (Artificial Intelligence or AI or A.I.) and ((ethic* and (guidelines or principles or "codes of conduct")), producing in total 256 results. The next step is "work", in which I identified the most interesting and contemporary publications that referred to weaknesses of implementations of the AI Ethical guidelines, leading to 7 main articles, focusing on contemporary publications from 2019 to 2021. These articles gain my attention due to the fact, that they were highlighting practical arguments and data on how the ethical guidelines of Artificial Intelligence seem insufficient in assisting the responsible and ethical development of AI. Finally, in the "wow" step, I am inclined to believe that it is wise to summarize my perspective that AI Ethical guidelines/ codes of conduct are not practically implemented and rephrase the problem on "What is the content of AI Ethical codes of conduct and why they are not practically implemented within the AI framework"? This problematization will continue in the next chapter, that I am going to explain the arguments of these scientific papers.

2.2. Ethical codes of conduct

The most philosophical inquiry here would have been: "Which kind of ethics are these Artificial Intelligence Ethical guidelines being referred to". Consequentialist ethics, deontological ethics, Aristotelian virtue ethics, more generical moral and ethical codes? How the ethical guidelines or principles are being connected with classical ethics or other ethical theories and how they are working on restraining a biased development of Artificial intelligence practices?

The current explosion in the development of AI applications has as a result the emergent call of applied ethics, ethical guidelines/codes that are going to control "metaphorically" and practically this unrestrained process. As an outcome, plenty of institutions, private and public organizations as well as academia are publishing in recent years a variety of ethical principles for the regulation of AI.

Nevertheless, as I realized during the aforementioned literature search, the main issue is that these principles are most of the time, unavailable to transform into technical codes and provide a concrete solution to ethical dilemmas. More specifically as Haggendorff describes in his Paper "The ethics of AI Ethics":

> "AI ethics – or ethics in general – lacks mechanisms to reinforce its own normative claims. Of course, the enforcement of ethical principles may involve reputational losses in the case of misconduct, or restrictions on memberships in certain professional bodies. Yet altogether, these mechanisms are rather weak and pose no eminent threat. Researchers, politicians, consultants, managers and activists have to deal with this essential weakness of ethics" *(*Haggendorff, 2020.:1)

However, even though there is this confusion on the usability of these ethical principles, whereas, ethics remain appealing to AI companies and institutions. As Haggendorff again describes:

> "When companies or research institutes formulate their own ethical guidelines, regularly incorporate ethical considerations into their public relations work, or adopt ethically motivated "self-commitments", efforts to create a truly binding legal framework are continuously discouraged…And even when more concrete laws concerning AI systems are demanded, as recently done by Google (Google 2019), these demands remain relatively vague and superficial." (Haggendorff, 2020.:1)

Therefore, Haggendorff here produces a different perspective on the development and reproduction of these AI Ethical principles, claiming that a practical legal ethical framework is constantly avoided or is being vague and abstract for the sake of industrial benefits. In the same notion, Mark Coeckelbergh expressed similar risks and questioned the practicality of ethical principles: "that ethics are used as a fig leaf that helps to ensure acceptability of the technology and economic gain but has no significant consequences for the development and use of the technologies" (Coeckelbergh as cited in Larsson, 2020:442). It can become evident, that ethical guidelines are being used as justification for certain actions.

Another perspective for the development of AI ethical principles and the difficulty of them is coming from Kearns and Roth who propose the following:

> "Kearns and Roth point to the fact that the speed, volume, and specificity at which algorithms are being developed surpasses the human timescale necessary to implement laws, regulations, and watchdog organizations. Their solution is to develop AI algorithms that internalize values of fairness, privacy, accuracy, and transparency; to ensure that algorithms are "better-behaved." (Kearns & Roth as cited in Royer 2020:6).

Moreover, another perspective that is similar to Haggendorff, Larsson and Royer, is coming from Bruce M. McLaren, in his article "Extensionally defining principles and cases in ethics: An AI model", in which he is pointing out the inability of practical implementation of AI Ethical codes:

> "While most people would agree that abstract principles such as these are reasonable and appropriate, it is difficult to apply them in real-world situations [21]. Since the principles contain open-textured terms and phrases… it is not possible for experts to define intermediate-level rules to cover all possible conditions to which the principles apply." (McLaren, 2019:146).

In addition to that, more publications underline the insufficiency of the practical implications of AI Ethical principles for different reasons. For instance, the paper "A critical perspective on guidelines for responsible and trustworthy artificial intelligence", from Banu Buruk et al, refers that ethical principles of AI are being developed mainly from computer scientists, engineers, and the voice of social scientists and philosophers is staying extremely low:

> "Techniques such as machine learning used in the development of AI technology are not framed from a value/principle based ethical perspective but rather, developed within the framework of economic logic. Since fast results are the most important criteria to the business world which seeks profit ethical evaluations are often ignored, and documents remain at the level of wishes, thus weakening the enforcement power of the ethics guidelines of AI." (Buruk et al, 2020:397)

In another example, in the paper "On conflicts between ethical and logical principles in artificial Intelligence", the importance of cultural differences and the difficulty of the objectification of what is good on a global scale is emphasized, "This is not just a matter of geographic, economic or cultural diversity: this lack of an objective notion of goodness is rooted in human nature." (D'Aguisto, 2019:897). Furthermore, in the paper "Superethics Instead of Superintelligence: Know Thyself, and Apply Science Accordingly", the impracticality of putting ethical codes into practice is being highlighted, given the fact that research is indeed attempting to comply with technical codes and identify ethical ways of developing the AI technology "The problem of homo sapiens is not to formulate ethical rules, it's to put them into practice. What we are still lacking is not the identification of the ethical values involved, but technology that supports us in promoting and endorsing those values." (Haselager & Mecacci, 2020:116). The same perspective is being featured in the article "Morality of Artificial intelligence", that recommendations are still far from practical implementations, despite research initiatives, there is not a legal context that has been applied, "Legislation of AI is still catching up to the progress made in research and practice, and there have not yet been any country level laws governing AI research specifically." (Luccioni & Belgio, 2020:21). Finally, the article "Machine Ethics, Allostery and Philosophical Anti-Dualism: Will AI Ever Make Ethically Autonomous Decisions?", is

pointed out the lack of ethical knowledge and education of the developers of AI Technology (Hauer, 2020).

Therefore, after the research on the different publications I made through the literature research, I am inclined to conclude that the practical implementation of AI ethical principles is lacking and is being insufficient due to the reasons that were mentioned above and you can moreover observe in the following table. Accordingly, I am presenting the following table, to review the different research perspectives on the weaknesses in the implementation of ethical principles in Artificial Intelligence.

Table 1, "Weaknesses on the practical implementation of AI Ethical Principles"

REFERENCE	STATEMENT	ARGUMENT
"Extensionally defining principles and cases in ethics: An AI model" Bruce M. Mclaren, 2019	"While most people would agree that abstract principles such as these are reasonable and appropriate, it is difficult to apply them in real-world situations [21]. Since the principles contain open-textured terms and phrases… it is not possible for experts to define intermediate-level rules to cover all possible conditions to which the principles apply." (McLaren, 2019).	Ethical Principles are being abstract, unable to be applied in practical "real world" situations.
A critical perspective on guidelines for responsible and trustworthy artificial intelligence Banu Buruk et al. 2020	"Techniques such as machine learning used in the development of AI technology are not framed from a value/principle based ethical perspective but rather, developed within the framework of economic logic. Since fast results are the most important criteria to the business world which seeks profit ethical evaluations are often ignored, and documents remain at the level of wishes, thus weakening the enforcement power of the ethics guidelines of AI." (Buruk et al, 2020:397)	Ethical principles remain on the wishes level because they do not possess a practical outcome for AI, which primarily focuses its development on economic logic.

On conflicts between ethical and logical principles in artificial Intelligence Giuseppe D'Acquisto, 2019	"This is not just a matter of geographic, economic or cultural diversity: this lack of an objective notion of goodness is rooted in human nature." (D'Aguisto, 2019:897).	An objective notion of what is good lacks in human nature, and it is difficult to be defined.
Superethics Instead of Superintelligence: Know Thyself, and Apply Science Accordingly Pim Haselager & Giulio Mecacci, 2020	"The problem of homo sapiens is not to formulate ethical rules, it's to put them into practice. What we are still lacking is not identification of the ethical values involved, but technology that supports us in promoting and endorsing those values." (Haselager & Mecacci, 2020:116).	The problem of humans, it to put the ethical rules into practice.
On the morality of Artificial Intelligence Alexandra Luccioni & Joshua Belgio, 2020	"Legislation of AI is still catching up to the progress made in research and practice, and there have not yet been any country level laws governing AI research specifically." (Luccioni & Belgio, 2020:21).	Research and practice of AI are developing so fast, thus the legal (ethical) rules are difficult to catch up with the progress.
Machine Ethics, Allostery and Philosophical Anti-Dualism: Will AI Ever Make Ethically Autonomous Decisions? Tomas Hauer, 2020	"Given the lack of the necessary ethical expertise among programmers working in machine learning, there is a great risk that AI researchers will build on incorrect ethical assumptions and develop their ethical approach to machines on precarious grounds" (Hauer, 2020)	Ethical expertise among developers is lacking, thus plenty of mistakes can happen in developing their ethical approaches.

The Ethics of AI Ethics: An Evaluation of Guidelines Thilo Haggendorff, 2020	"AI ethics – or ethics in general – lacks mechanisms to reinforce its own normative claims. Of course, the enforcement of ethical principles may involve reputational losses in the case of misconduct, or restrictions on memberships in certain professional bodies. Yet altogether, these mechanisms are rather weak and pose no eminent threat. Researchers, politicians, consultants, managers and activists have to deal with this essential weakness of ethics" (Haggendorff, 2020.:1)	Ethical principles lack the technology to support their normative claims. The guidelines seem weak at this level.

Finally, it is evident that ethics are lacking practical implementation in Artificial Intelligence systems, for a variety of reasons mentioned in the table above, such as "they lack the technology to support their normative claims or that the software developers are lacking ethical expertise or that they seem only rules of compliance" among others. Additionally, to strengthen my argument I want to outline a practical experiment by McNamara, Smith and Murphy Hill, that was an attempt to assess if ethical guidelines can be implemented. Therefore, they gathered 168 software developers (engineers and experts), and they provided them with eleven software-related ethical decision scenarios to assess if they ethical guidelines that were using may affect the decision making in six vignettes. "The results were disappointing: no statistically significant difference in the responses for any vignette was found across individuals who did and did not see the code of ethics, either for students or for professionals." (McNamara, Smith, and Murphy-Hill 2018:4 as cited in Haggendorff, 2020:6). Concluding with this last argument I am inclined to believe that the practical implementation of Ethics in AI, is insufficient and there is a necessity to repair the content of the ethical framework, but first we need to understand it.

2.3. Problem Statement

Finalizing my problem analysis, I am inclined to believe that, to "assist" the confrontation of impracticalities in the ethics of AI, I am obliged firstly, to identify the different codes of conduct of ethics in AI (gather them in a database) and understand their content, which is a unique methodology, comparing to the theoretical outlines that I found through my literature search. Secondly, I will create an innovative methodological perspective, by bringing notions of techno-anthropology and more specifically digital methods to analyze the ethical principles of AI, since digital methods offer the opportunity to see at many codes of conduct and analyze (digital text analysis) the ethical principles that they recommend. Finally, by understanding the ethical principles that seem impractical for the AI framework, outline how a combination of Aristotelian (Virtue) Ethics, Foucauldian Ethics and Techno-Anthropology

can provide unique perspectives on the development of the AI Ethical framework in practice.

Concluding my problematization, I am presenting the declarative problem statement as follows:

Due to the lack of practical implementation of the ethical principles in the Artificial Intelligence systems, this techno-anthropological research seeks to understand this phenomenon and how we can assist in the development of the AI Ethical framework by answering the following:

1. What is the content of the AI ethical codes of conduct?
2. How techno-anthropological methods can offer a different perspective on the understanding of AI Ethical principles?
3. How Aristotelian, Foucauldian Ethics and Techno-Anthropology can assist in the development of AI Ethical principles?

3. Methodology

In this chapter, I am going to unveil the methodological background that I used for this thesis project, which consisted of two main components, the theoretical framework and the computational methods I used for the data collection and data/text analysis. The term "Method" can be traced back to Aristotelian philosophy as the tradition and series of actions focused to reach a specific target "the route that leads to a target" (During as cited in Markopoulos, 2018:51). Accordingly, my main intention here is to follow a specific route of techno-anthropological methodological approaches, to identify and analyze the AI Ethical codes of conduct and how we can assist their development.

3.1. Theoretical Framing

In the first sub-chapter of the methodology, I am going to outline the theoretical framework of the thesis project that has been utilized as the main ground for the analytical approach of this research. Thus, I am starting with a theoretical outline of the main ethical theories (descriptive, normative, deontological, consequentialist, virtue, meta-ethics, and applied ethics), based on the book "An Introduction to Ethics in Robotics and AI" of Christoph Bartneck, Christoph Lutge, Alan Wagner and Sean Welsh, underlining perspectives from "Episteme, technology and philosophical reflection" of I.N. Markopoulos. Moving on, I am delving into theoretical perspectives of Foucault in "Conditions of our Freedom: Foucault, Organization and Ethics" by Andrew Crane, David Knights, and Ken Starkey, in which I am analyzing ethical notions that are highlighted in Foucault's theory and correspondingly Foucault's concept "technologies of the self".

Subsequently, I am delineating in "Aristotelian Ethics" and summarizing Aristotelian views about Virtue ethics and more specifically intellectual and ethical virtues with focusing on the concept of "phronesis" and the concept of responsibility, based in the book of Christof Ratt "Introduction to Aristotle". The target of the theoretical framework is to offer to the reader, a background of the classical ethical theories that the concept of Ethical codes of conduct of Artificial Intelligence has been based on. These theories will be used in combination with data/text analysis, in the analytical part of this research project, to decipher the meaning and the content of the most popular ethical principles that are being used on the Artificial Intelligence systems. Additionally, the Aristotelian and Foucauldian ethics will be used in the discussion part to outline recommendations for an innovative AI Ethical framework.

3.2. Data collection

In this chapter, I am outlining the methodologies that I used for my data collection and the practices I used to analyze them. Firstly, my method was to generate a database with different AI Ethical principles, gathering codes of conduct from different sectors, such as academia, white papers, and various institutions. Attempting that, I went across different websites and search engines, to reach the "Algorithmic Watch website – AI Ethics Guidelines Global

Observatory",[4] in which one can find various codes of conduct with recommendations on AI Ethical principles. These codes of conduct were in form of white papers, developed from governmental and international agencies, files published from technological private organizations and scientific papers developed by academia. It is important here, -thus we can avoid any confusion- to restate again the dichotomy, between AI Ethical codes of conduct (files/papers/research) and AI Ethical principles which are the AI Ethical values that are being referred to, inside these codes of conduct.

To collect the various codes of conduct on that website, I developed a python coding script (See Appendix B), managing to download approximately all the files, referred to AI Ethical principles. For a few papers, that did not have a document edition (pdf), but only an online version, I manually downloaded them. Afterward, I composed a database with one-hundred and sixty-two different codes of conduct in an excel sheet (See Appendix C), with the assistance of a scraper software (Instant Data Scraper). In this excel sheet, I manually added the different ethical values/principles that every document stated (Guideline 1-14), the title, the author, the year, the country that the research has published and the link of the document. For the process of the manual addition of every guideline, I had to identify in every one of those codes of conduct, the chapter that was referring to the ethical guidelines/principles for Artificial Intelligence such as transparency, responsibility, safety, security, etc. and add them to the excel database as values.

Moving on, in the analysis part, I am using a python script (See Appendix B), to calculate the number of occurrences of every value, to identify the most popular principles that are being referred into the global AI Ethics guidelines inventory and the less popular principles, to detect which are the trends and if certain ethical principles/values are being overlooked.

3.3. Digital Methods

In the final part of the methodology, I will outline how I utilized the collective data and how I analyzed them with the help of digital methods and more specifically digital text analysis. As I am reflecting in the previous chapters of methodology, my narration is starting with the data collection, the data observation and classification which have as a result the analysis of the most popular ethical principles of Artificial Intelligence with the help of ethical classical theories.

To delve more into the content of these specific ethical principles of AI, I am using with the help of my education in Techno-Anthropology, a combination of theory and practice, a theoretical narration in ethical theories combined with a digital text analysis that offers me a unique opportunity to pinpoint definitions of the principles. The procedure I am using to manage this digital method approach is the following. Firstly, I am gathering a small number of definitions about the specific AI Ethical value/principle (e.g., Transparency), from the codes of conduct on the Ethical AI database. Secondly, I am creating a file that isolates these mere definitions (See Appendix A). Finally, with the help of "Voyant tools" which is an

[4] Algorithmic Watch – "AI Ethics Guidelines Global Inventory": https://inventory.algorithmwatch.org/. April 2020. Accessed October 20, 2021.

application for digital text analysis, I can identify the words that are being used the most in each definition, how each value is connected to each other and useful information regarding the text formation. Finally, in combination with the classical ethical theories, this digital methodology can offer interesting perspectives that I will highlight in the analysis part of this research project.

4. Theoretical Framework

Starting this theoretical outline, I want to focus on outlining classical ethics methodologies that should be taken under consideration in the Artificial Intelligence different sectors design. In the following chapters, I am going to describe different ethical theories that are used -or not- as the basis of contemporary ethical guidelines. However, before highlighting the classical ethical theories, I want to delve into the meaning of ethos in technology.

4.1. Ethics in Technology

For Tuchel, "Technology has a general meaning, regarding methodologies, objects, and programs based on innovation and improving constantly for the gratification of personal and social needs. Due to normative functions, they are reaching targets and changing the society and the world." (Tuchel as cited Markopoulos, 2018:21). Going way back to Ancient Greece, the world technology is composed of two parts, "techne" and "logos". Techne for Aristotle means the virtue of technical rationality (Børsen, 2019:14) while "logos[5]" in ancient Greek, means target. Hence, the meaning is that technology is developing to reaching a target and optimize some social or individual needs. As a result, technology is always progressing in a socio-technical concept and it is essential to respect specific guidelines to develop in a certain social context. When I am referring to ethics of technology (in this particular thesis, ethics of Artificial Intelligence), I am not only discussing rules or restrictions that the society or the individual has to comply with but instead an action that is evolving "metaphorically" deeper than compliance and laws and becoming the way of life.

Before moving into the discussion of the ethical theories, it is important to state the difference between Ethics and compliance. Unfortunately, it is likely problematic that most of the ethical guidelines for Artificial Intelligence are being used by technological organizations as a matter of compliance, as a series of rules and restrictions that they have to respect to avoid fines or social turmoil by the masses. Cigrefs' paper is accurately describing the separation between these domains:

> "Compliance is all about operating in accordance with a standard or a law –something external which has authority. It is therefore everyone's responsibility to abide by the law or face sanctions. Ethics, on the other hand, is a personal or collective approach (at company level, for example) which entails setting guidelines for oneself. This approach is based on values or principles that can guide one's actions. Ethics is an act of empowerment (and not only responsibility), engagement and integrity." (Syntec Numerique & Cigref, 2018: 7)

Ethics are not just moral codes that individuals or organizations have to obey to avoid financial or social agitation, neither are doctrines that are going to unveil what is right and

[5] Britannica "logos". https://www.britannica.com/topic/logos . N.d. Accessed June 3, 2021

wrong. Instead, they are offering to individuals the ability to criticize, observe and identify an appropriate way of confronting moral matters that are going to appear in someone's life.

Ethos

Going a little back in time we can witness for the first time the term Ethics in Ancient Greece, coming from Aristotle with the word "Ethos", which was originally referred to habit, custom, the continuous practice of moral action, whereas Cicero, some centuries ahead translated the term into Latin and "mores" from which we are identifying the contemporary concept of Morality (Cicero 44bs as cited in Bartneck et al, 2021:17). Whereas the term "Ethics" is often getting the same meaning as morality, it should be wise to distinguish them on most occasions. As defined in the book "An introduction of ethics and robotics in AI by Bartneck et al,

> "Morality refers to a complex set of rules, values and norms that determine or are supposed to determine people's actions, whereas ethics refers to the theory of morality. It could also be said that ethics is concerned more with principles, general judgements and norms than with subjective or personal judgements and values refers generally to a complex set of rules, values and norms." (Barneck et al, 2021:17)

In the following Centuries, plenty of philosophers, sociologists, and other scientists, such as Immanuel Kant, Friedrich Nietzsche, Michael Foucault, and others developed their ethical theories, which I am going to describe in the following chapters without any particular order. The reason I am referring to these theories is to offer the reader the opportunity of having a general overview of the different ethical theories that have been developed through time.

Descriptive Ethics

Descriptive Ethics is the category that I will start with. They are dealing with the explanation of normative systems, like experimental economics or moral psychology. An example that Guth is using to formulate the meaning of Descriptive ethics is the following: "experimental results exhibit certain features of moral intuitions of people: studies using the "ultimatum game" show that many people have certain intuitions about fairness and are willing to sacrifice profits for these intuitions" (Guth et al 1982 as cited in Barneck et al, 2021:18). It is a category of ethics that can be utilized as an important input for normative ethics, especially when the normative evaluation of specific actions seems impossible, or the principles of evaluation are inadequate. Speaking about normativity, allow me to introduce Normative Ethics to the next chapter of ethical theories.

Normative Ethics

The **s**econd category I want to refer to, is Normative ethics, the ethics of "good" and "bad", of "morally correct actions" or "morally wrong actions." Normative Ethics are getting in the conversation when someone wants to identify, what is morally correct or morally wrong. For instance, the practice of stealing something is generally an action that is assumed to be morally wrong. Accordingly, "Normative ethics is usually not regarded as a matter of subjectivity, but of general validity. Stealing is wrong for everybody. Different types of normative ethics make judgments about actions on the basis of different considerations."

(Barneck et al, 2021:19). Normative Ethics, theoretically are judging and evaluating the ethical behavior of the individual, concerning a specific action. The main two categories that the Normative Ethics are divided into, are the Deontological ethics and Consequentialist Ethics that are the two following categories that I will describe in this thesis project.

Deontological Ethics

Deontological Ethics is the first sub-category of Normative Ethics that I am going to briefly describe. Deontological Ethics is the ethical theory that evaluates the ethical "correctness" of specific actions, on the "basis of characteristics that affect the action itself (Barneck et al, 2021:19). For instance, an attribute like this may be the purpose with which an action is carried out or the compatibility with a specific formal concept. The consequences of the action are not being the basis of the evaluation and judgment here, although they are taken under consideration. Again, the terminology is coming from the ancient Greek "Deon", which can be translated as duty or obligation, hence Deontology can be translated as duty Ethics (Barneck et al, 2021:19).

There are plenty of practical examples that deontological ethics can utilize their normative ability. For instance, a case in Greece that is probably a controversial issue to other countries, similarly, is the ability of rich people to provide donations for charity, while they are getting away from taxes. In Greece (possibly in other countries too), it is common knowledge that when a rich person is offering a big amount of donation, is escaping the tax system[6]. Hence, even though providing goods to the people who are in need is theoretically a moral and "good" action, there is a large percentage of people, that tend to believe that this hypothetically morally good action, cultivates different intentions which are not morally acceptable. A common example that is used in moral dilemmas is correspondingly the famous "trolley problem[7]", as well as autonomous cars are other examples that deontological ethics can be applied to.

Immanuel Kant is one of the most important figures of deontological ethics and the responsible of some of the most important citations about them. As referred to the introduction of ethics in robotics and AI, Kant argues that,

> "An action is only obligatory if it satisfies the "categorical imperative". There are many different wordings of the categorical imperative, which is best understood as a way of determining ethically permissible types of behavior. The most frequently cited version states, "Act only according to that maxim you can at the same time will as a universal law without contradiction." (Barneck et al, 2021:20)

In conclusion, I may define Deontological Ethics as the Ethical theory that examines if action is complying with appropriate duty, or the suitable norm. Deontological Ethics are a basis for ethical dilemmas and one of the most important theories, that brought back Ethics into the

[6] Europa "Tax Edu": https://europa.eu/taxedu/young_el . June 3, 2021. Accessed June 3, 2021

[7] The trolley problem: https://www.theguardian.com/science/head-quarters/2016/dec/12/the-trolley-problem-would-you-kill-oneperson-to-save-many-others . December 2016. Accessed June 3, 2021

practical discussion and principles formation. Their opposing theory, but at the same time likely the theory that completes them is the next one.

Consequentialist Ethics

Consequentialist Ethics is the second sub-category of Normative Ethics, that I am going to briefly describe in this thesis project. Instead of judging the intentions of the specific actions, Consequentialist Ethics are evaluating the ethical correctness of the action or a norm by their potential consequences. For instance, in the aforementioned example with the donations in Greece, from the moment that the donation of the rich individual, saved some lives or provided some shelter, food to people in need, this is a morally good action, an ethical behavior, because here only the outcome is being evaluated, not the initial intentions. However, the outcome of actions can be abstract. For instance, in self-driving cars, in an autonomous technology such as this, the outcome is not based solely on the intentions of the driver or the machine and even if the driver -in a case of an accident- wants to avoid injuring someone else, this can be a potential outcome, thus this is a morally wrong practice and unethical behavior of the driver, according to deontological ethics, even if their intentions were not to harm someone. I am going to discuss more paradigms such as this, later in this thesis project, in the analysis part, where I am going to refer to the classical ethical theories.

In conclusion, Consequentialist Ethics are the ethics that are focusing on the effects of specific practices and their short or long-term negative or positive outcomes (Syntec Numerique & Cingref, 2018). Utilitarianism, which is a subsector of consequentialism, states that always we should aim to the result that is going to save the most people in a case of an accident for instance.

Virtue Ethics

The next theoretical concept that I am going to introduce is Virtue Ethics. To outline them, I have to return to Ancient Greece and the famous Greek Philosophers Plato and Aristotle. Plato developed the concept of the virtues (wisdom, justice, fortitude, and temperance) and Aristotle expanded it later, adding intellectual virtues, while "The classical view on virtues held that acting on their basis was equally good for the person acting and for the persons affected by their actions". (Barneck et al, 2021:20). In Syntec & Cigref document, virtue ethics are defined like the ethical theory that:

> "Describes the moral character of action according to the accompanying virtue. People talk of courageous, just, and generous acts, for example. In this conception, it is the courses of action and moral attributes of the person that are most important." (Syntec Numerique & Cingref, 2018:9).

I am going to return to virtue ethics and Aristotelian virtues, later in this theoretical analysis. Generally, there is a controversy on if virtue ethics are still applicable in modern societies, although the controversy is expanding in ethics in general. I am going to return to this statement later in this thesis project.

Applied Ethics

The last category, that I am going to briefly refer to in this chapter is Applied Ethics. As it can easily be interpreted from their name "Applied Ethics", are the Ethics that are willing to be applied -into different situations-. Thus, ethical theories, considerations, that are seeking ground to be applied, practices to show the true applications of ethics. They seem similar to normative ethics and meta-ethics, but instead, they are referring to more concrete fields where ethical judgments can be applied, like biotechnology (bioethics), medicine, business (business ethics) (Barneck et al, 2021). Applied ethics, are being used differently on every occasion and argue that there is not a universal code, but external powers and social constructions can alter and adjust their application. A suitable definition is the following one, "Applied Ethics puts normative ethics into practice, by comparing a concrete situation with principles derived from various schools of normative ethics. Ethical dilemmas are always resolved using applied ethics." (Syntec Numerique & Cingref, 2018:9). Applied Ethics, should be a more comprehensive study, delving into the ethical considerations, that developed from the classical Ethical theories and adjusting them into practice.

4.2. Foucauldian Ethics

A theoretical mindset that I am inclined to believe, that will be proven inspiring for the analysis of the contemporary Ethical principles in Artificial Intelligence, is the French philosopher, sociologist, and anthropologist Michael Foucault and his insightful perceptions about ethical theory, because of his theoretical outlines about ethics and power structures. Foucault is famous especially for his theoretical perspectives on power structures and agency, and concepts such as the "Panopticon", but it is extremely interesting to outline here notions of his general theoretical ideologies, combined with ethical perceptions. In order, to explore the Ethical framework that AI is being implemented and the power structures that enhance it, it is essential here, to present notions from the Foucauldian Ethics.

As referred to in the paper "Conditions of our Freedom: Foucault, Organization and Ethics" by Andrew Crane, David Knights, and Ken Starkey, for Foucault: "ethics is a practical concept, used to denote the possibilities of individual agency . . . rather than following a religiously-based norm, or acting in accordance with some Kantian transcendental imperative" (Styhre, 2001:799 as cited in Crane et al, 2008:304). Here Foucault, underlines the capability of ethical theories to be applied and not the generic view of evaluating "good and bad" actions. For Best and Kellner "Ethics concerns not so much moral norms as it does free choice" (Best & Kellner as cited in Crane et al, 2008:303). Therefore, it is essential to highlight here the perspective of Foucault, about ethics being a practical concept and not normativity based on religion, while free choice (autonomy) is a mandatory ethical framework.

Speaking about free choice and freedom in general, Foucault's belief -based on his emphasis on disciplinary forces and structures- is that "freedom is necessarily limited and can never be absolute, thus a society without restrictions is inconceivable" (Foucault, 1997: 148 as cited in

Crane et al, 2008:303). The restriction for Foucault is something that cannot be avoided; thus freedom has limits and these limits are emerging from society itself. As Foucault describes:

> "I don't believe there can be a society without relations of power, if you understand them as means by which individuals try to conduct, to determine the behavior of others. The problem is not of trying to dissolve them in the utopia of a perfectly transparent communication [as suggested by Habermas], but to give one's self the rules of law, the techniques of management, and also the ethics, the ethos, the practice of self, which would allow these games of power to be played with a minimum of domination." (Foucault, 1994: 18 as cited in Krane et al, 2008:304)

According to Foucault, in a society one has to participate actively in the law restrictions, management techniques, and generally the power relations, thus there can be a balance and most importantly a confrontation on the domination of the powerful agencies. One can argue here, that this is a concerning matter due to the fact that, the technological unrestricted development of Artificial Intelligence, in which the current rules, restrictions, or laws are being, solely a matter of a specific group of people (the dominating actors) and not a concept that is going under discussion with different societal groups (the ones who need to be aware of the imbalance and confront the power structures).

Moving deeper to the theoretical perception of Foucault concerning ethics, I have to provide his definition of ethical behavior here. According to Foucault ethical behavior is the result of:

> "[A] process in which the individual delimits that part of himself that will form the object of his moral practice, defines his position relative to the precept he will follow, and decides on a certain mode of being that will serve as his moral goal. And this requires him to act upon himself, to monitor, test, improve, and transform himself.... A moral action tends toward its own accomplishment; but it also aims beyond the latter, to the establishing of a moral conduct that commits an individual, not only to other actions always in conformity with values and rules, but to a certain mode of being, a mode of being characteristic of the ethical subject." (Foucault, 1985a: 28, as cited in Crane et al, 2008:305)

Hence, regarding Foucault, ethical behavior (of a human), is a process of testing, monitoring, and improving themselves, a series of actions that will have as a result the establishment of a "moral conduct".

Technologies of the self

If we get back to the Kantian notions of deontological ethics, one can argue that Foucault is opposing Kantian imperatives and rationality. However Foucault is not entirely challenging Kant in the individual's duty to exercise self-control, both for self-improvement and the benefit of society (Foucault, 1985a; Best & Kellner, 1991; Danaher et al., 2000, as cited in Crane et al, 2008:304) Foucault's ideologies on individual agency, power structures and freedom (free choice) are bringing some perspectives on the matter of ethical behavior and these perspectives are summarized into Foucault's notion of "Technologies of the self", that according to Foucault are "those intentional and voluntary actions by which men not only set themselves rules of conduct but also seek to transform themselves, to change themselves in their singular being" (Foucault, 1985a: 10-11 as cited in Crane et al, 2008:305)

This ideology is going back to Ancient Greece and Aristotelian views, which I am going to outline in the next chapter. The examinations of these perspectives led Foucault to extensive studies in Ancient Greece and Rome, where concepts such as “the care of oneself, toward definite objectives such as retiring into oneself, reaching oneself, living with oneself, being sufficient to oneself, profiting by and enjoying oneself” were the core of the individual’s ethical behavior (Foucault, 1988a, as cited in Crane et al, 2008). In addition to that, Rabinow describes that “through forms of self-definition and self-constraint, people train themselves to become ethical persons. From this view flowed the idea that the self is not given to us, but that "we have to create ourselves as a work of art" (Rabinow, 1984: 351 as cited in Crane et al, 2008:305).

Hence, Foucault delineates his theoretical outcome on Ethics as an ability that we can use to potentially transform ourselves for the “common good” of society as well as the individuals. According to Foucault: “The ethical task is to challenge oneself as one is, or finds oneself, and to "take oneself as an object of a complex and difficult elaboration…Thus modernity does not 'liberate man in his own being'; it compels him to face the task of producing himself” (Foucault ethics 7-8 as cited in Crane et al, 2008:305) In the same notion, Foucault rejects the idea of an authentic, absolute self but rather champions an "ethics of creativity" as opposed to an "ethics of authenticity" (Owen, 1994: 201-02, as cited in Crane et al, 2008:305). More specifically concerning technologies of the self, Ransom adds that:

> “As such, technologies of the self do not have a desired end state, but are an ongoing set of reflexive practices that work and rework the self in relation to disciplinary power. Developing technologies of the self therefore involves developing oneself into someone who is more aware of the possible effects of disciplinary procedures and, therefore, better able to resist them (Ransom, 1997: 139, as cited in Krane et al, 2008:306).

Therefore, “technologies of the self”, is a theoretical concept that outlines how an individual may be aware and prepared to confront power domination or technological (AI) domination in our framework, by getting to understand the different hierarchies and structures of power in a situation, to adjust and resist every time, with a separate practice in this network of power. Hence, here it becomes evident the purpose of Foucault, to outline that the awareness and the demystification of power structures are essential for humans.

A great notion for the theory of the application of ethical theories in Artificial Intelligence and general in algorithms is coming from Rorty who underlines that "God has provided no algorithms for resolving tough moral dilemmas, and neither have the great secular philosophers," (Rorty (2006: 371) as cited in Crane et al, 2008:306). Additionally, Foucault states that "we should not waste our time searching in vain for universals. Where universals are said to exist, or where people tacitly assume they exist, universals must be questioned" (Flyvberg, 1998:222 as cited in Crane et al, 2008:307). The discussion for universal rules and universal ethical codes or applications of ethical principles into algorithms and Artificial Intelligence is something, that as Foucault and other philosophers argued is a waste of time. I will return in this, later into the analysis and discussion part of this thesis.

From the previous paragraphs, we can understand that Foucault has a different view about ethical discussion and freedom, putting in the conversation the power structures and forms of domination.

> "Foucault acts ethically in taking the risk of giving us choice: once people have been given the vital knowledge of how forms of power have acted upon and constructed them, then they are "left to make up their own minds, to choose, in the light of this, their own existence" (1988b: 50). This insistence upon giving the other free ethical choice is the closest Foucault ever comes to laying down a moral code" (Krane et al 2008:308).

An approach similar to this, but still different is the one of Habermas who sees ethics primarily residing in "the institutionalization of... procedures and conditions of communication" in democratic decision-making processes. In a business context, this suggests attention to democratic control on the public use of corporate power (Scherer & Palazzo, 2007 as cited in Crane et al, 2008:308). As Andrew Crane, David Knights, and Ken Starkey also mention, "in relation to pragmatism, a refusal to acknowledge the importance of power in enacting non rule-based systems of business ethics would be seen as incomplete and naive from the perspective of Foucault's ethics." (Crane et al, 2008:308)

> "The care of the self … implies complex relationships with others insofar as this ethos of freedom is also a way of caring for others... Ethos also implies a relationship with others, insofar as the care of the self enables one to occupy his rightful position in the city, the community, or interpersonal relationships, whether as a magistrate or a friend. Thus, the problem of relationships with others is present throughout the development of the care for the self." (Foucault, 1997: 287 as cited in Crane et al, 2008:311).

The aforementioned statement is offering me the opportunity to move safely to the next chapter, that I am going to unveil more about the word ethos and the Aristotelian view about morals, virtues, and political, social relationships embedded into the conversation about ethics.

4.3. Aristotelian Ethics

As Aristotle defines, "philosophy is the research of the causes and principles ("περι τα πρώτα αίτια και τας αρχάς")" (Markopoulos, 2018:22), thus ethics are researching these principles. This chapter is a much-needed examination of Virtue Ethics, developed mainly by the Ancient Greek philosopher Aristotle. The reason, that I am presenting notions from the Aristotelian/Virtue Ethics here, is to accomplish a general discussion that I will outline in the analytical and discussion part (Chapter 5-6) of this thesis project, to explore and attempt to develop and reshape the contemporary AI Ethical guidelines.

Most of the classical economical theories, which are starting with Plato and Aristotle are combine three sectors: the "Oikos", the "individual", and the "polis". The development of the "individual" is a concept connected with the other two forms, the "Oikos" and the "polis", thus an individual is growing with internal influences, coming from the house (Oikos) and the society, city (polis) is living in. Therefore, when we are referring to ethical behaviors in

artificial intelligence systems, we should bear in mind all the fundamental principles that construct an individual's character in a society (Larsson:2020). More specifically in Aristotelian ethics, for one to accomplish the ultimate welfare and "eudaimonia" one has to intend not only for the individual welfare but instead for the prosperity of their society.

For Aristotle, applied philosophy includes ethics as well as the political philosophy and the "praxis" theory (action theory). Aristotle is focusing on the term politics in his attempt to define ethical behavior, since an Ancient Greek ideology was that the "Real political leader" is the one who can unleash the potential of individuals and make them better and for this matter has to possess ethical abilities (Plato, Gorgias as cited in Rapp 2012:19). Why the applied philosophy and applied ethics accordingly are important? Because regarding Aristotle, the ultimate target of a person is "Eudaimonia". As Aristotle describes, one can reach the "Eudaimonia", if one already knows their ultimate target "As an archer who has to look the target, to reach the target" (Ethica Nichomacheia 1094a, as cited in Rapp:2012:19).

However, are applied ethics and philosophy the answers to what is morally good and what is morally wrong? Aristotle describes that firstly, is impossible to deal with moral dilemmas with the same accuracy as geometry can solve a mathematical problem. It would be unfair to ask Ethics for correctness like this and then perceive ethics as something insufficient if it cannot solve an issue. Secondly, applied philosophy fails to decide on what is "good" or "wrong" for a person in a specific situation. As Aristotle underlines, this is a concept that applied rationality, the virtue of "phronesis", is dealing with these matters. Finally, applied ethics and philosophy is not something that can be randomly taught to someone, instead, it is an ability that one can possess with experience and practical dilemmas (Rapp, 2012:20).

Aristotelian Ethics are based on the theory of pursuit of happiness and "good life", or "eudaimonia" as Aristotle defines, the ultimate happiness and good living, "As the one who adores horses, appreciate the contact with horses, the same way the one who is an admirer of the virtues, get appreciation when accomplishing virtuous actions (Rapp, 2012:27). On this statement, and the general theory of "eudaimonia" one can argue that the ultimate target of an individual should be accomplished without caring about the impacts to others. In contrast to this opinion, Aristotle highlights that firstly, the quest of "eudaimonia" is opposing selfishness at the expense of others and secondly, recognizes that the prosperity and welfare of an individual are connected with the prosperity and welfare of the society (Rapp, 2012).

Ethical and Intellectual Virtues

An essential concept for Aristotelian Ethics is the concept of virtues. Aristotle distinguished the virtues into two clusters, the "intellectual" virtues, and the "ethical" virtues. Firstly, intellectual virtues are including knowledge (episteme), wisdom, "techne" and the practical rationality "phronesis", which is the practical operation of decisions that are affecting the ultimate target of "eudaimonia". To achieve "phronesis" one has to possess general knowledge and additionally get experience from taking decisions considering moral dilemmas, alone or with the advice of the teacher and the laws (Rapp, 2012).

Secondly, the ethical virtues are the ones that are related to emotions and desires, such as generosity, bravery, justice, prudence, gentleness, and magnanimity and these ethical virtues

are not innate to a human being but are being acquired by a repeated routine into moral actions and humans have the predetermination on obtaining these abilities (Rapp, 2012). Individuals have a predisposition to gaining these talents, even though experience, repetition, and regular exercise are all essential for cultivating and practicing skillsets. Humans have the predetermination "pefykosi" of acquiring the described skillset, according to Aristotle, since they are not inborn (Maor et al, 2019). As an analogy for this point, he uses a natural phenomenon:

> "The stone will always move downwards because it obeys the natural law of gravity, which is constant and unchanging. The fire will always move upwards due to the natural property of the hot gases, which is also constant and unchanged. Thus, it follows from the foregoing that natural laws do not change, no matter how much one tries. On the other hand, a man with his actions and choices can change his behaviour, cultivate and develop some qualities of his character." (Athanasopoulos, 2013-2014)

Therefore, intellectual virtues are a matter mainly of education whereas ethical virtues need education and repetition (Rapp, 2012). However, the most important contributor in the definition of ethical virtues is the notion that virtues constitute the middle point between two negative sides "mesotis" (mediocrity). For instance, bravery forms the "mesotis" between excessive fear and excessive risk.

Aristotelian Responsibility

A fundamental inquiry for Aristotelian philosophy and in general a question that is popular amongst the circles of those who are researching the development of technology and especially artificial intelligence is "who is responsible?" and when an action is forgivable. Thus, when an action is intentional? A traditional answer could be that an action is intentional when is caused by free will. For Aristotle, "an unintentional action is the one which is caused because of oppression or ignorance" (Ethica Nichomacheia 1109, as cited in Rapp, 2012:40).

Aristotle is also referring to a category between intentional and unintentional actions and he is mentioning an example of how a sailor has to unleash the cargo in the sea due to natural phenomena (storm) and blackmailing (Ethica Nichomacheia 1110, as cited in Rapp, 2012:41). This sailor can choose to oppose the blackmailing or not to throw the cargo, which of course will have as a result, the destruction of the ship. On the other side, one can be considered as responsible for an action, when there is no ignorance and the action has a focused intention. For Aristotle, free will is divided into physical desires and emotional desires (Peri psychis 432 b, as cited in Rapp, 2012:45). I will discuss more about responsibility on Artificial Intelligence, in the analysis part of this thesis project.

5. Analysis

In this part of this research project, I am unveiling the AI Ethical Guidelines database that I generated from my data collection. The purpose of this analysis is to identify the various AI Ethical principles that have been proposed while deciphering their content and meaning throughout combined perspectives from classical ethical theories and the digital text analysis approach.

5.1. AI Ethical codes of conduct

The target of this chapter is the analysis of the ethical principles that were gathered from different codes of conduct, mainly from the website "Algorithmic Watch - AI Ethics Guidelines Global Inventory". To analyze these principles, I collected one hundred and sixty-two different codes of conduct, across the world that were suggesting different notions of ethical principles for Artificial Intelligence. These codes were either white papers and governmental research, or individual researchers/academia papers, or industry papers that were published mainly from big-tech companies. Hence, as I am referring to in chapter 3, my data collection methodology was to collect the codes of conduct referring to AI Ethical principle, by a python script (See Appendix B) and identify manually in every code, the appropriate context about AI Ethical principles and gather all the different principles in an Excel database (See Appendix C), with one-hundred and sixty-two different results. For the full AI Ethical guidelines database, with all the information and results, you can check Appendix C.

	Link	Country	Year	Institution	Principle 1
2	https://www.aomrc.org.uk/wp-content/uploads/2019/01/Artificial_intelligence_in_healthcare_0119.pdf	United Kingdom	2019	Academia	Patient Safety
3	https://www.accenture.com/t20160629T012639Z__w__/us-en/_acnmedia/PDF-24/Accenture-Universal-Principles-Data-Ethics.pdf	United States	2016	Private Sector	Privacy and Securit
4	https://www.accenture.com/gb-en/company-responsible-ai-robotics	United Kingdom	2018	Private Sector	Decision making ar
5	https://www8.cao.go.jp/cstp/tyousakai/ai/summary/aisociety_en.pdf	Japan	2017	Government	Awareness (emoti
6	https://ai-white-paper.readthedocs.io/en/latest/	Italy	2018	Government	Responsibility
7	https://ainowinstitute.org/AI_Now_2018_Report.pdf	United States	2018	Academia	Fairness / Justice
8	https://www.torontodeclaration.org/	United Kingdom	2018	Civil Society	Human rights
9	https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf	United States	2017	Industry Association	Awareness
10	http://www.eismd.eu/wp-content/uploads/2019/02/Ethical-Framework-for-a-Good-AI-Society.pdf	European Union	2018	Civil Society	Justice
11	https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf	Australia	2019	Government	Privacy Protection
12	https://www.baai.ac.cn/blog/beijing-ai-principles	China	2019	Government	Good no Harm
13	https://algorules.org/en/home	Germany	2019	Civil Society	Responsibility
14	https://www.bitkom.org/sites/default/files/file/import/150901-Bitkom-Positionspapier-Big-Data-Leitlinien.pdf	Germany	2015	Industry Association	
15	https://www.bitkom.org/Bitkom/Publikationen/Empfehlungen-fuer-den-verantwortlichen-Einsatz-von-KI-und-automatisierten-Entscheidungen-Corporate-Digital-Responsibility-and-Decision-Making.html	Germany	2018	Industry Association	
16	https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=4	Germany	2019	Government	
17	https://ki-verband.de/ki-guetesiegel-ai-made-in-germany	Germany	2019	Industry Association	
18	https://cdt.org/wp-content/uploads/2018/09/Digital-Decisions-Library-Printer-Friendly-as-of-20180927.pdf	United States		Civil Society	Fairness
19	https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-ai-alliance-drafts-self-discipline-joint-pledge/	China	2019	Voluntary	Security
20	https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-expert-group-offers-governance-principles-responsible-ai/	China	2018	Government	Harmony
21	https://www.cigionline.org/publications/toward-g20-framework-artificial-intelligence-workplace	Canada	2018	Civil Society	Privacy
22	https://www.cigref.fr/wp/wp-content/uploads/2019/02/Cigref-Syntec-Digital-Ethics-Guide-for-Professionals-of-Digital-Age-2018-October-EN.pdf	France	2018	Industry Association	Data protection
23	https://www.cssf.lu/fileadmin/files/Publications/Rapports_ponctuels/CSSF_White_Paper_Artificial_Intelligence_201218.pdf	Luxembourg	2018	Government	Data protection
24	https://rm.coe.int/2018-lignes-directrices-sur-l-intelligence-artificielle-et-la-protecti/168098e1b7	European Union	2018	Government	Proportionality
25	https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf	United States	2018	Civil Society	Non Discriminatior
26	https://dataethics.eu/data-ethics-principles/	Denmark	2017	Civil Society	Human centered A
27	https://hippocrate.tech/	France		Civil Society	Transparency
28	https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf	Norway	2018	Government	Data protection
29	https://deepmind.com/safety-and-ethics	United States		Private Sector	
30	https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology	United Kingdom	2019	Government	
31	https://www.telekom.com/de/konzern/digitale-verantwortung/details/ki-leitlinien-der-telekom-523904	Germany	2018	Private Sector	Responsibility
32	https://www.dgb.de/themen/++co++4f242f08-18a7-11e9-b2c1-52540088cada	Germany	2019	Civil Society	
33	https://www.digicatapult.org.uk/projects/machine-intelligence-garage-ethics-framework/	United Kingdom	2020	Private Sector	Benefits
34	https://www.smartdubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf?sfvrsn=d4184f8d_6	United Arab Emirates	2019	Government	Ethical codes
35	https://em.dk/media/12190/dataethics-v2.pdf?utm_campaign=Background&	Denmark	2018	Government	Fairness
36	https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/	United Kingdom	2010	Government	
37	https://www.ethikbeirat-hrtech.de/wp-content/uploads/2019/09/Ethikbeirat_und_Richtlinien_Konsultationsfassung_final.pdf	Germany	2019	Private Sector	

Figure 3, AI Ethical Guidelines Database, (See Appendix C)

By analyzing the data in the spreadsheet above, my target is to identify which are the principles that are being referred, to the most inside the documents, which are the principles that are referring the less, and how these AI Ethical values/principles are connected or not,

with classical ethical theories, while identifying which potential principles that are coming from classical ethical theories are being excluded or referring rarely.

Before delving into the calculation, it would be wise to find out, who is responsible for the development of these guidelines, more specifically which country. Therefore, according to statistics[8] coming from Algorithmic Watch, the US is the country publishing the most documents regarding AI Ethical guidelines with forty-four. International documents (UN, IEEE, G20, G7, UNESCO) are in second place with twenty-five publications. Afterward, Germany has twenty publications and UK nineteen, whereas documents from the European Union are eights, as well as France. Moreover, Canada has six publications, and China has four, the same as Japan. Other European countries that published more than one, are Finland (three), Netherlands (three), Denmark (two), Italy (two), and Switzerland (two). In Asia, Singapore and South Korea also published two papers accordingly, while in Oceania, there are only two publications in total, one from Australia and one from New Zealand. Finally, in Africa, there is only one code of conduct on recommendations for AI Ethical principles, published in South Africa. You can see in the following data visualization, the publications per country.

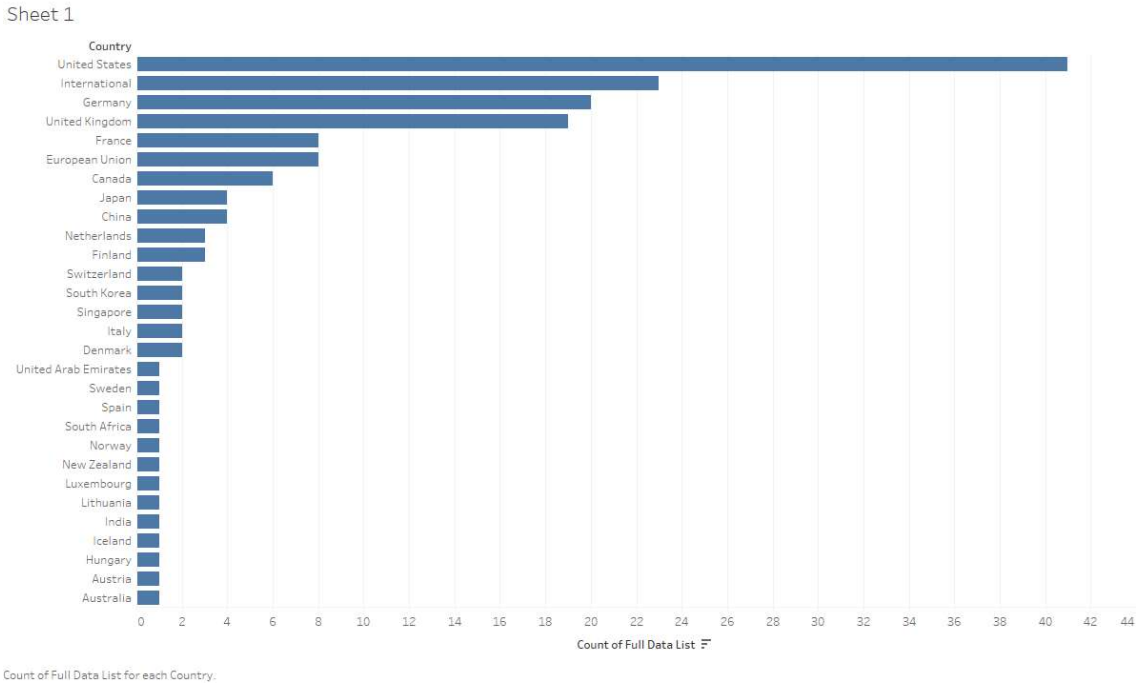


Figure 4, Publications of AI Ethical codes of conduct per country, made by Tableau

I am presenting this number, comparing continents, due to the fact, that this dichotomy and the huge development of publications in Europe (in total seventy-three) and Northern America (in total fifty), whereas Asia has only fourteen in total, Oceania two, Africa one and Southern America zero, represents consequently the inequality globally and the smaller number of opportunities. It is possible that governments in Africa and Southern America are not publishing more documents on the AI, not because they are not interested in the

[8] Algorithmic Watch -AI Ethics Guidelines Global Inventory/ Region-Location- https://inventory.algorithmwatch.org/ April 2020. Accessed May 30, 2020.

technology, but instead because there are likely other issues that are necessary to deal with right now, whereas not having the opportunity to deal with this technology, due to the fact that Artificial Intelligence development may be expensive. Furthermore, it is an obvious fact, that these AI codes of conduct are being influenced and referring to the West, eliminating from the conversation non-western societies, leading to an exclusion of the social and cultural impact in the AI Ethics conversation.

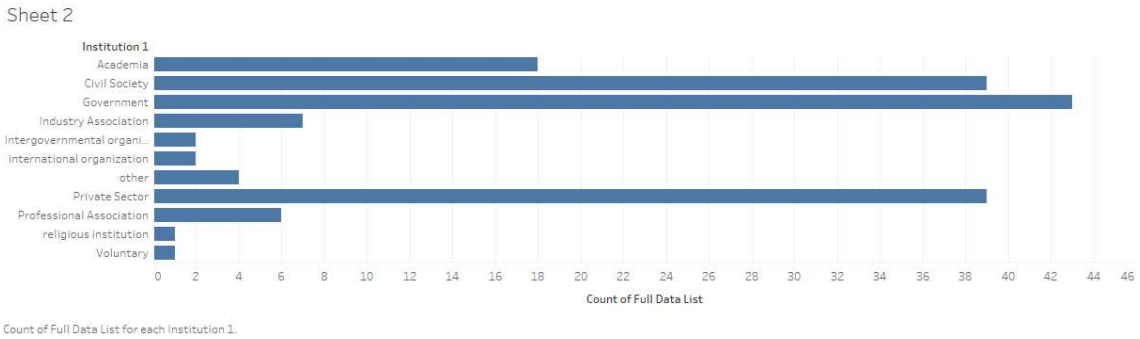
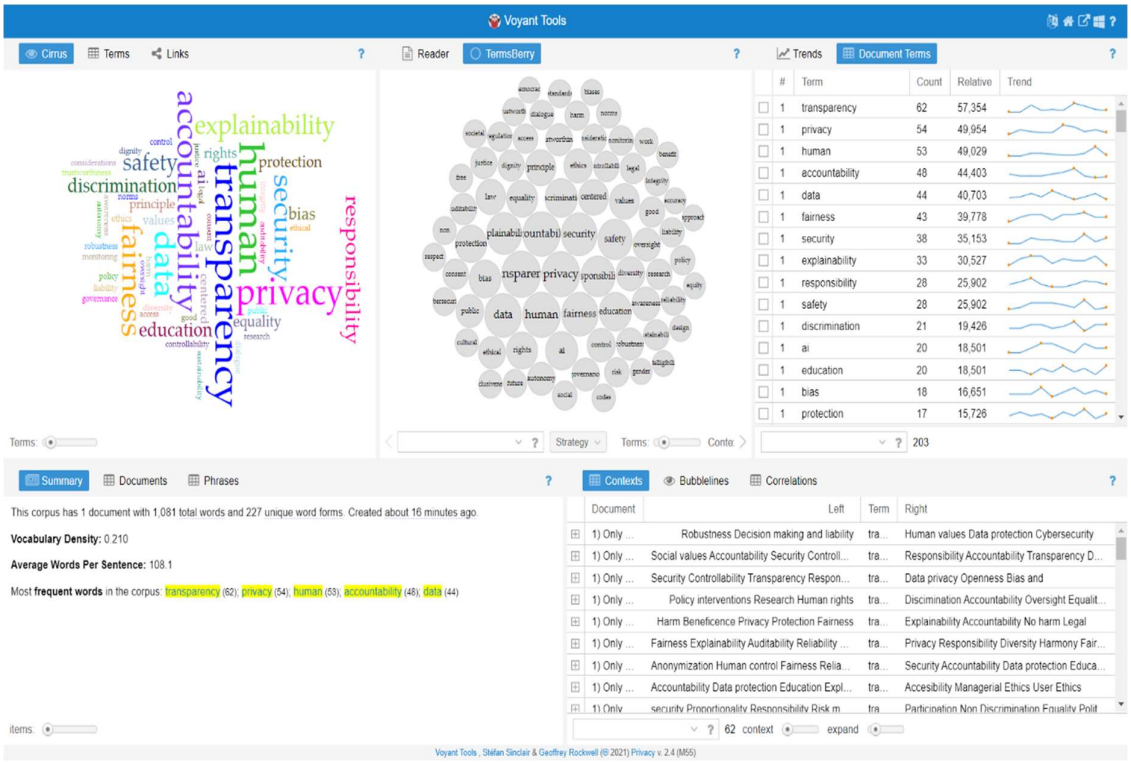


Figure 5, Publications of AI Ethical codes of conduct per various sectors made by Tableau

Additionally, in the data visualization above, you can witness the different sectors, that have been published the AI Ethical codes of conduct, rising questions about the power structures of Western Societies and the imbalance between Academia and private sectors. In chapter 6.4. of discussion, I will delve more into this matter.

5.2. Deciphering the AI Ethical Principles

The main part of the data collected here is to analyze the Ethics of AI Ethical guidelines. Therefore, with the help of "Voyant Tools" we can witness the most referred AI Ethical principles in the database and accordingly the less referred. Accordingly, you can check the Appendix C, for the calculation.



#	Term	Count	Relative	Trend
1	transparency	62	57,354	
1	privacy	54	49,954	
1	human	53	49,029	
1	accountability	48	44,403	
1	data	44	40,703	
1	fairness	43	39,778	
1	security	38	35,153	
1	explainability	33	30,527	
1	responsibility	28	25,902	
1	safety	28	25,902	
1	discrimination	21	19,426	
1	ai	20	18,501	
1	education	20	18,501	
1	bias	18	16,651	
1	protection	17	15,726	

Document	Left	Term	Right
1) Only ...	Robustness Decision making and liability	tra...	Human values Data protection Cybersecurity
1) Only ...	Social values Accountability Security Controll...	tra...	Responsibility Accountability Transparency D...
1) Only ...	Security Controllability Transparency Respon...	tra...	Data privacy Openness Bias and
1) Only ...	Policy interventions Research Human rights	tra...	Discimination Accountability Oversight Equalit...
1) Only ...	Harm Beneficence Privacy Protection Fairness	tra...	Explainability Accountability No harm Legal
1) Only ...	Fairness Explainability Auditability Reliability ...	tra...	Privacy Responsibility Diversity Harmony Fair...
1) Only ...	Anonymization Human control Fairness Relia...	tra...	Security Accountability Data protection Educa...
1) Only ...	Accountability Data protection Education Expl...	tra...	Accesibility Managerial Ethics User Ethics

Figure 6, Calculation of AI Ethical guidelines Database, See Appendix C

According to the data calculation above, the most popular value/principle in the AI Ethical guidelines database was the principle of transparency, which appeared in sixty-four from one-hundred and sixty-two documents. Secondly, the principle of "privacy" underlined in fifty-four from one-hundred and sixty-two documents, followed by the principle of "accountability" in forty-eight from one-hundred documents. Afterward, we can witness, the principle of "fairness", written in forty-three texts as well as the principles of "security", thirty-eight times, responsibility twenty-eight times, and safety twenty-eight times. Education is similarly high with twenty appearances in the one hundred and sixty-two documents. Moreover, discrimination, bias, and equality are likewise mentioned twenty, eighteen and sixteen times accordingly in the texts of this database. Finally, principles such as autonomy and Justice are mentioned seven times in the database.

If you look closely at the data calculation above, one will argue about the terms "human" and "data", which are coming third with fifty-three mentions and fifth with forty-four mentions accordingly. I am not referring to these terms, in the description above, because these are just words that are part of different AI ethical principles such as "Human-Centered AI" or "Human rights" or "Human dignity", and "Data surveillance", or "Data privacy", or "Data protection. However, one can argue here, the importance of the word human in the AI ethics discussion. Indeed, the word human is being mentioned consequently, thus the ethical principle of "Humanity" should not be excluded from the conversation and the analysis below.

On the other side, in the database above we can identify terms that are not being used with the same intensity. Terms such as "social and cultural" are solely three times mentioned in the document, "precautionary principle" only one, whereas ethical principles such as phronesis, zero times.

In the next chapters, I will define the most popular principles of the aforementioned database and the ones that are missing from the conversation. To outline these principles, I will use the theoretical background from the frameworks that I delineated in chapter four in combination with techno-anthropological ethics and digital text analysis.

Transparency

I will start this outline of AI Ethical guidelines, with the principle of "transparency", which was the most popular principle, being mentioned in the majority of the codes of conduct. If we return, to the classical ethical theories of chapter 4, one can argue that there is not a single mention of the principle of "transparency" in classical Ethical theories. Thus, what is transparency as an ethical AI principle? As Tom Børsen argues "Transparency requires one to operate in such a way, that it is easy for others to see what actions are performed and what decisions are made…Transparency implies openness, communication, and accountability". (Børsen, Ethical 2019). More importantly, in the article "On the Governance of AI Ethics", transparency is being mentioned as a "pro-ethical condition", for enabling other ethical principles or practices (Larsson, 2020:445).

Hence, I would attempt to describe transparency, as a principle behind the ethical codes, that is enabling and forcing other principles to emerge. A principle that has to be a part of the ethical framework to protect mainly human rights. Furthermore, to delve deeper into the meaning of transparency as an ethical principle of Artificial Intelligence, I narrowed down several definitions of the term as being introduced in the database of AI Ethical guidelines and attempted a text analysis with the help of "Voyant tools" software that you can observe below.

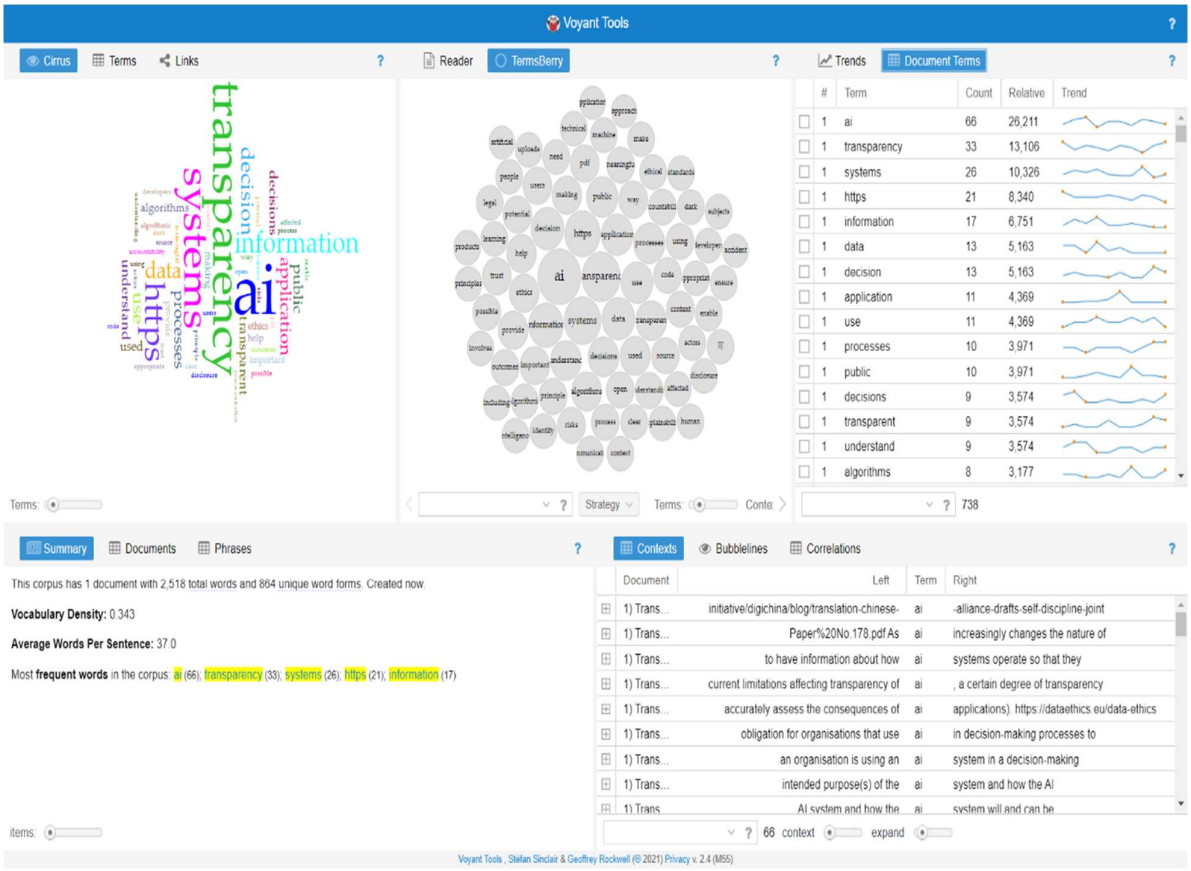


Figure 7, Transparency principle, digital text analysis

By observing this text analysis, one can argue that the definitions of transparency as an ethical principle of Artificial Intelligence, are focusing on the decision(s) making, the understanding of the processes and the transparent information about its processes that the Artificial Intelligence systems, must provide (Keywords: decision(s) {22}, information {17}, data {13}, processes {10}, understand {9}). Accordingly, Leslie defines transparency in his research "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector." (Code of conduct in the AI Ethical guidelines database), as a metaphor for opening the black box of the AI system that assists the trust, safety/security and openness, explainability, justifications procedures (Leslie, 2019). Thus, indeed transparency can be defined as a pro-ethical principle that assures that an artificial intelligence system is being developed with the notion to be transparent, open and explainable to its users and ensures that the other principles (justice, accountability, autonomy, safety), will also be present in the development.

Fairness / Justice

Moving on, I am clustering these two terms because generally, they are sharing the same meaning, translation. Fairness is the fifth most popular term in the guidelines' database, with forty-three mentions, whereas the term Justice is correspondingly being mentioned seven

times. The meaning is quite the same and the most contemporary word of fairness is following the route of the most traditional principle of justice. As referred to in the book "An introduction to ethics of Robotics and AI",

> "The principle of justice states that AI shall act in a just and unbiased way. Justice is often illustrated by a statue of the Roman Goddess Justitia. Frequently, she is depicted with a sword, scales, and a blindfold). The blindfold represents impartiality. The scales represent the weighing of evidence. The sword represents punishment." (Bartneck et al, 2021:33)

At the same time, in Techno-Anthropological Ethics, Tom Børsen describes that Justice can have two definitions. Firstly, actions to generate the biggest benefit to the least advantaged members of society and secondly, that everybody must be treated equally according to their merit and effort. (Børsen, 2019). Finally, as discussed in chapter 4.3., Aristotle ranks Justice in the Ethical virtues, that can be possessed through repetition and exercise, therefore if we follow the inductive methodology of Aristotle, the ethical principle of justice may be owned from human beings, only through repetition and exercise, such the same repetition can create ethical Artificial intelligence systems, if one follows the Aristotelian reasoning.

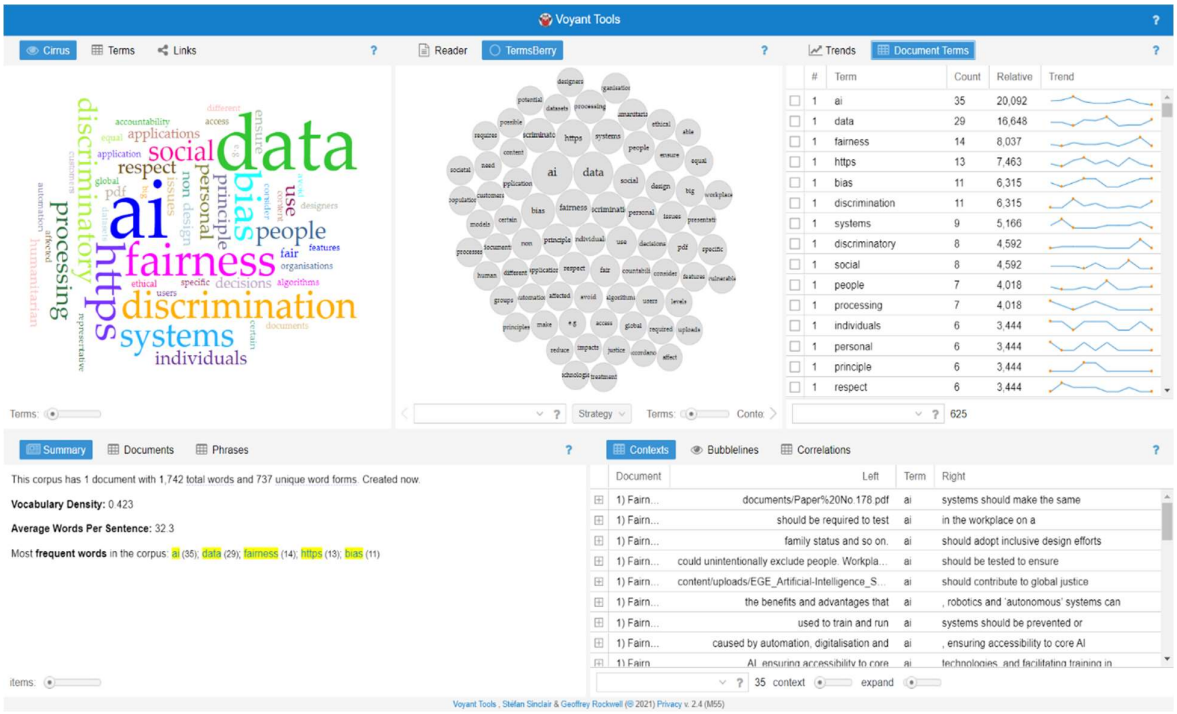


Figure 8, Fairness principle, digital text analysis

Observing the text analysis visualization above, that accomplished through definitions of fairness principle from the Artificial Intelligence Ethical guidelines database, one can argue that fairness principles is referring mostly to bias and discrimination incidents and seek to confront the bias in the development of Artificial Intelligence systems, assisting the notions of respect and equality (keywords: bias {11}, discrimination* {19}, respect {6}, data {29}). Accordingly, Telia company (codes of conduct database), defined the principles of fairness,

as a method to confront biases of discrimination and inequality (Telia company, 2019). The fairness/justice principle is the only one of the most acknowledged (in the database) principles, that can be identified back on virtue ethics.

Furthermore, a high number of biased AI cases, regarding discrimination and inequality have been brought to the spotlight lately by "Awful AI[9]" curated list in Github. In the following table, you can explore various cases of discrimination and unfairness in AI.

Table 2, "Discrimination and Unfairness Biased AI cases"

Reference	Bias
AI-based gaydar https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph (Levin, 2017)	Artificial intelligence can accurately guess whether people are gay or straight based on photos of their faces, according to new research that suggests machines can have significantly better "gaydar" than humans.
Racist chat bots https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter (Hunt, 2016)	Microsoft chatbot called Tay spent a day learning from Twitter and began spouting antisemitic messages.
Depixelizer https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias (Vincent, 2020)	An algorithm that transforms a low-resolution image into a depixelized one, always transforms Obama into a white person due to bias.
Sexist Recruiting https://www.telegraph.co.uk/technology/2018/10/10/amazon-scraps-sexist-ai-recruiting-tool-showed-bias-against/ (Cook, 2018)	AI-based recruiting tools such as HireVue, PredictiveHire, or an Amazon internal software, scans various features such as video or voice data of job applicants and their CVs to tell whether they're worth hiring. In the case of Amazon, the algorithm quickly taught itself to prefer male candidates over female ones, penalizing CVs that included the word "women's," such as "women's chess club captain."
Gender detection from names	Genderify was a biased service that promised to identify someone's gender by analyzing

[9] https://github.com/daviddao/awful-ai/blob/master/README.md. "Awful AI". December 30, 2020. November 9th, 2021.

https://www.theverge.com/2020/7/29/21346310/ai-service-gender-verification-identification-genderify (Vincent, 2020)	their name, email address, or username with the help of AI. According to Genderify, Meghan Smith is a woman, but Dr. Meghan Smith is a man.
Persecuting ethnic minorities https://www.theguardian.com/news/2019/apr/11/china-hi-tech-war-on-muslim-minority-xinjiang-uighurs-surveillance-face-recognition (Byler, 2019	Chinese start-ups have built algorithms that allow the government of the People's Republic of China to automatically track Uyghur people. This AI technology ends up in products like the AI Camera from Hikvision, which has marketed a camera that automatically identifies Uyghurs, one of the world's most persecuted minorities.

Accountability

The next value that I am deciphering is the third most popular with forty-eight mentions in the codes of conducts database, the principle of "Accountability". One can argue, that the principle of explainability can be clustered with the principles of accountability, because according to Floridi et al describing, Explainability or Explicability can have the meaning of both intelligibility and accountability, and they are focusing as principles on affecting AI systems to make sense on the certain way of "how AI systems are making certain decisions" (Floridi et al, as cited in Bartneck et al, 2021:36). However, I am inclined to believe that the meaning of the principles of explainability is closer to the meaning of transparency, thus I will focus solely on the principle of accountability here.

However, how the principle of accountability is connected to classical ethical theories? According to the book "Towards a code of Ethics for Artificial Intelligence", Plato has explained briefly the value of accountability:

"…A related aspect is accountability to others. A corollary of Plato's views on knowledge and government is that, in governing those under them, the 'noble lie' could be justified to keep the "hoi polloi" in order. I take it that a view is abhorrent in any democratic society. It goes without saying that you can't claim to be adequately addressing ethical questions, if you refuse to explain yourself to rightly interested parties." (Boddington, 2017:20)

In conclusion, the AI Ethical principle of accountability is being utilized with the target of constructing the AI technology to be precise and explainable to users or other interested parties, similar to how techno-anthropology is suggesting the unboxing, the "demystification" of the Black box, the decoding of Artificial Intelligence structure in concrete pieces, explicable to individuals. Accordingly, Plato has already -centuries ago- underlined the

importance of accountability as a fundamental of an ethical framework. Additionally, in the text analysis visualization below, created by definitions of accountability from the AI Ethical principles database, you can witness the connection of accountability with responsibility (Keywords: Responsib* {14}, developers {6}, design {6}, users {6}).

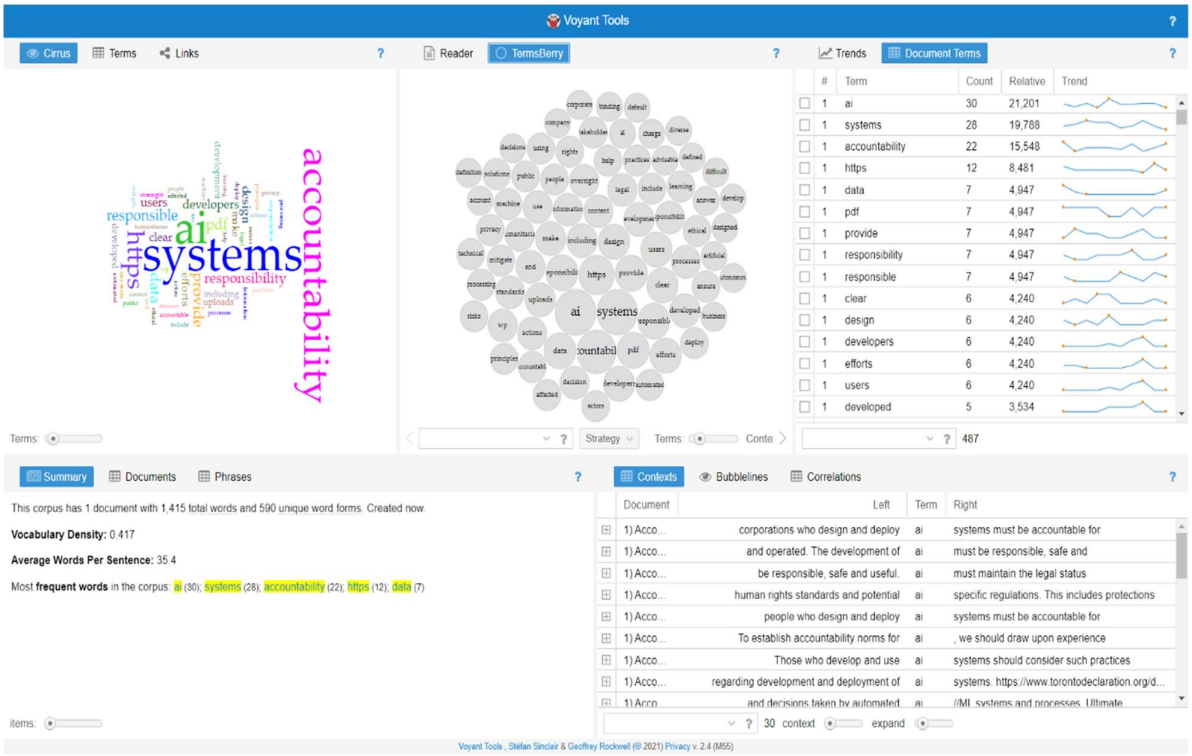


Figure 9, Accountability principle, digital text analysis

Regarding to IBE, accountability is the principle that has to ensure that there is always a line of responsibility in business actions, to justify who has to answer the consequences (IBE, 2018). Accordingly, in the principle of accountability, we are witnessing the attempt for the normativity of ethics with notions from the consequentialist ethics, due to the fact that the consequences are the means, to decide, who is the responsible one, depending on their work during the development/design of the AI system (More about responsibility, will be discussed in the next subchapter). Nevertheless, one can argue that the principle of accountability fails in certain cases to activate its normativity, as we can observe in chapter 1.2, in the Amazon Alexa case, when in the situation of murder, Amazon denies being accountable for their data and prefers to hold a neutral position, withdrawing any kind of responsibility in that case (Wang, 2017).

Responsibility

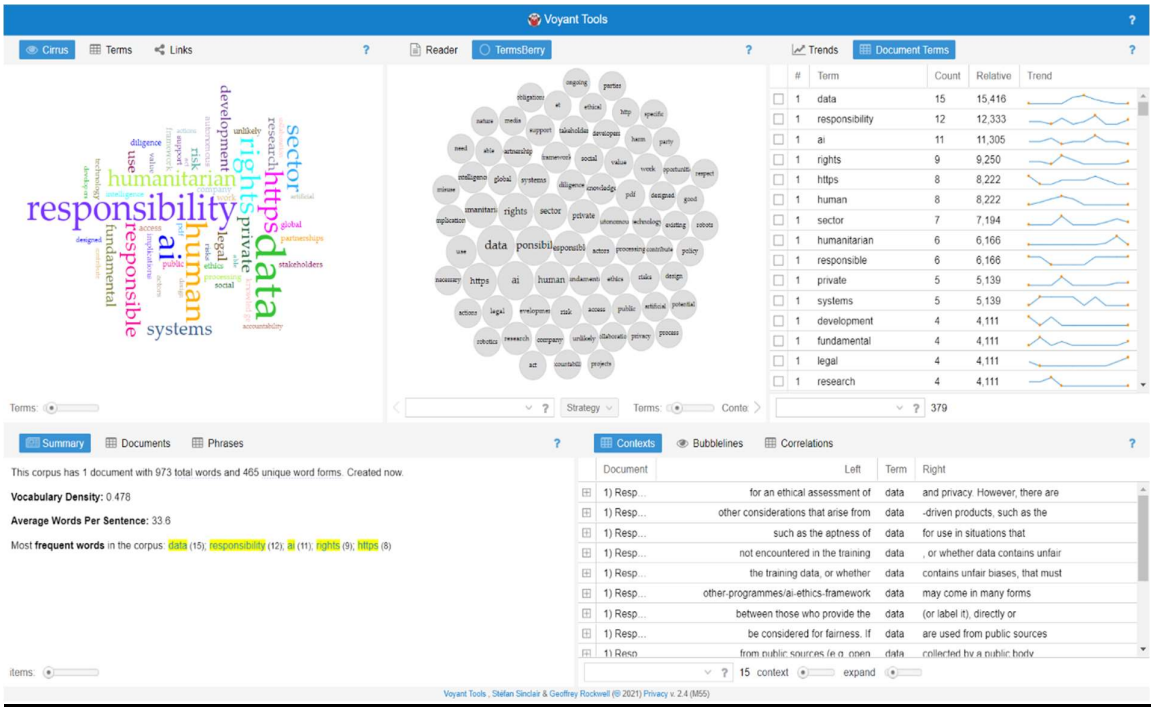


Figure 10, Responsibility principle, digital text analysis

In the text analysis visualization above, one can argue that responsibility as a principle is not only referring to the artificial intelligence systems to be responsible but moreover to the importance of human responsibility in the design and the development of the AI technology as well as the legal responsibilities (Keywords: human* {14}, development {4}, legal {4}). Thus, the responsibility one can argue should always be on humans and never on technology.

Responsibility is also a value mentioned in a high number (twenty-eight) of AI Ethical guidelines and one of the most popular principles in the discussion for AI ethical design and development. The main question when someone is developing an Artificial Intelligence application and this goes wrong is "Who is responsible for the bias". Is it the software developer, the engineer, is it the Artificial Intelligence system itself, due to the ignorance of the designer? As we can observe in digital text analysis, the normative perspective of the ethical principle of responsibility is focusing on identifying the human (developer, engineer, worker), as accountable and responsible for a negative consequence of the procedure without getting under consideration, external influences.

However, in virtue ethics, for Aristotle, the important question is when an action is intentional and when it is unintentional. Nevertheless, he is also underlining the existence of a third category of responsibility, which is in the middle of intentional and unintentional actions, using the example of the sailor to argue for his position.

> "..how a sailor has to unleash the cargo in the sea due to natural phenomena (storm) and blackmailing. This sailor, can choose to oppose the blackmailing or not to throw the cargo, which of course will have as a result, the destruction of the ship. On the other side, one can be considered as the responsible for an action, when there is not

ignorance and the action has a focused intention" (Ethica Nichomacheia 1110, as cited in Rapp, 2012:41).

Hence, different circumstances are essential here, to identify who has the responsibility. Firstly, the intention of the designer, secondly, one should take into consideration the circumstances under which, the engineer developed the AI system or which external factors influenced their decision, and finally, it is essential to identify the different scenarios on how a bias can unveil and take precautions in any case. Responsibility and precautions will be discussed later again in this thesis project.

Humanity

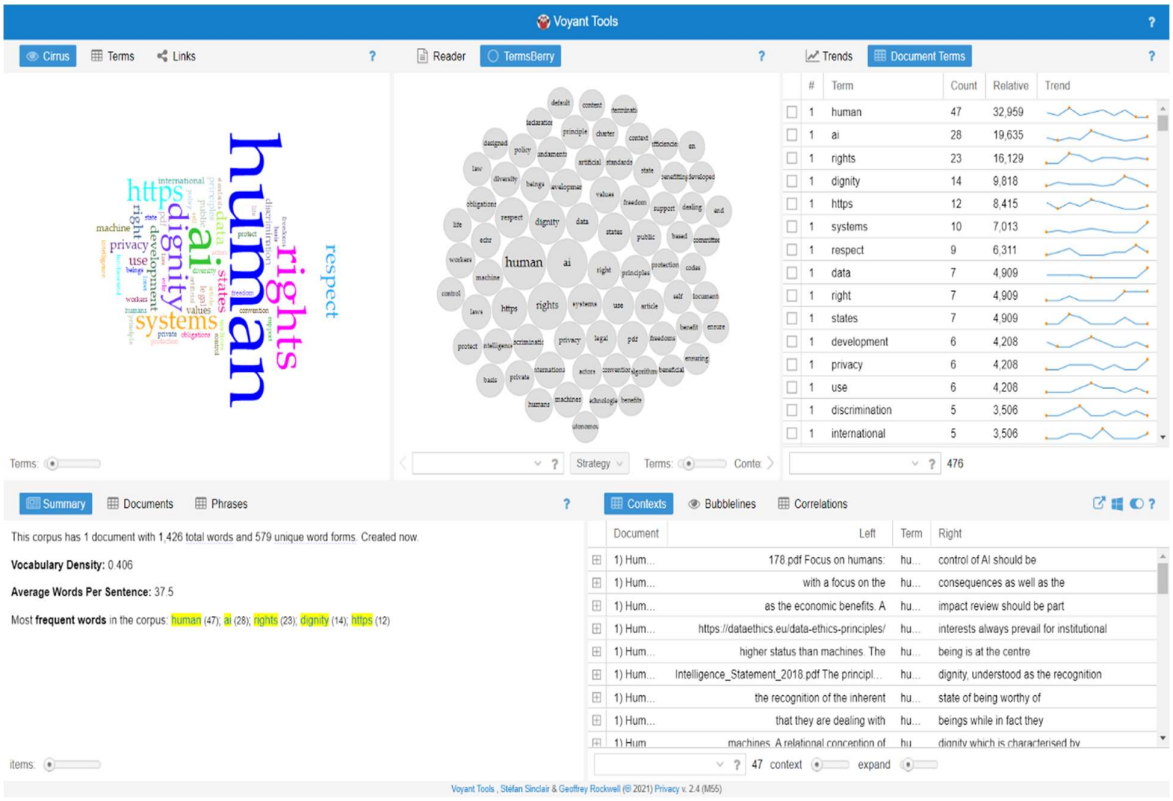


Figure 11, Humanity principle, digital text analysis

A central concept to the AI Ethical guidelines is the principle of Humanity, either as "Human rights", or "Human Dignity", or "Human-centered AI", or just humanity. In the text analysis visualization above, one can observe that most of the definitions for the principle of humanity are referring to human rights and then to human dignity and respect (Keywords: dignity {14}, right* {30}, respect {9}). In that sense, the AI principle of humanity is connecting with fairness and justice (human rights), whereas it is moreover targeting the importance of human dignity and respecting each other. Human dignity as AI principle should be understood as

> "…the recognition of the inherent human state of being worthy of respect, must not be violated by 'autonomous' technologies… 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." (EGE, 2018:18)

Hence, Human is an essential contributor to Artificial Intelligence ethical development, either as a contributor or as a user and the principle of humanity should have a concrete meaning. According to the book "An Introduction to ethics of Robotics and AI"

> "The principle of humanity tells us we should respect the ends (goals, wishes) of other persons. We should not treat other human beings as "mere means" alone. Rather we must consider the ends they have as well as our own." (Bartneck et al, 2021:32).

Furthermore, techno-anthropological ethics and Tom Børsen, described two attributes that are similar to the ethical principle of humanity, firstly, "compassion", "as the ethical value of helping another person who is suffering…Compassion is related to a vulnerability which obliges a person to help another who cannot withstand the hostile environment" and secondly "humility" which is the "Ethical value that is the anti-thesis of committing Hybris" (Børsen, 2015:87). In the same notion, Aristotle is underlining the importance, of acting not only for the benefit of yourself but for the benefit of the whole society that you are a member of, to reach the supreme target of every ethical being, the "eudaimonia". Hence, humanity is an ethical principle that needs to be re-constructed and being more concrete as AI ethical principle and not referred to as "Human-centered AI" or "Human Dignity", instead of as a principle that will lead the AI systems, to not overlook certain groups, on the benefit of someone.

Beneficence

In a high number of the documents in the database of AI Ethical guidelines, there was being mentioned the term "Beneficence", "No harm but good", "More benefits than harmful", etc. However, the real question here is how one can calculate the good and the bad? The answer can be identified, back in the utilitarian ethics and the "categorical imperative" of Immanuel Kant, when the consequences of the action are the ones that define if the result is to the benefit of individuals or it is more harmful than good, or more specifically "morally good" or "morally bad".

According to the idea of Beneficence, AI can benefit citizens. This is a general bioethics concept that states that the benefits of treatment must outweigh the risks. (Zalta 2003 as cited Bartneck et al, 2021:30). Hence, in the Artificial Intelligence systems, the morality of every programming action can be defined throughout the outcome and the benefit or harm to people according to the principle of beneficence, which seems like an ethical principle, born from the normativity of deontological and consequentialist ethical perceptions. However, the main weakness of the theoretical concept of "Beneficence" is that the outcome of "good" or "bad" is likely to be subjective in most cases, thus for instance, the utilities of AI may be beneficial and profitable in innovative workplaces, but on the other side, plenty of individuals cannot adjust in the new way of working, that has as a result, to lose their jobs. In this case, the company is getting profitable, but the worker is being harmed from the AI. Thus, how one

can calculate the Beneficence? It is a principle that underlines the social and cultural necessity in the ethical framework, regarding the different cultural notions on what is "morally good and bad".

Privacy / Autonomy

Privacy is the principle that comes second in the AI Ethical guidelines database, with 54 mentions across the different codes of conduct. Autonomy is a principle that is not mentioned a lot of times in the database, more specifically, only seven. Nevertheless, undoubtedly, they are, identical values and surely privacy/autonomy is a principle embedded into Artificial intelligence systems, due to the fact, that in a different scenario there is no moral responsibility from the person. According to the book "An Introduction to ethics of robotics and AI" Autonomy is referring to the following definition:

> "When a person individually, makes decisions. The principle of Autonomy states that AI shall respect people's goals and wishes. Moral philosophy is directly linked to autonomy. If a person does not have autonomy or free will, then it can be argued that this person does not have moral responsibility either (Bartneck et al, 2021:30).

The essentiality of autonomy hides, into the meaning of being responsible for your actions, which we are going to discuss later in this thesis project. Nevertheless, a keyword here is the term free will. Going back to Foucault here, and his statement about free will, we can witness how the principle of individual autonomy of AI systems is essential to be combined with free will,

> "Foucault acts ethically in taking the risk of giving us choice: once people have been given the vital knowledge of how forms of power have acted upon and constructed them, then they are "left to make up their own minds, to choose, in the light of this, their own existence" (1988b: 50). This insistence upon giving the other free ethical choice is the closest Foucault ever comes to laying down a moral code" (Krane et al, 2008:307-308)

The aforementioned statement is deriving from the "technologies of the self", a statement of Foucault, in which he is underlining the importance of having the own choice and the responsibility of actions and existence in general. Hence, the main message here is not only accusing the Artificial Intelligence bias but instead taking into consideration the responsibility of the individual's free will, human bias and autonomy. Accordingly, in the conference of G7 toward "AI Network Society on 2017" in Japan, the privacy principle highlighted the importance of how the developers should take into consideration the privacy of users or third parties as well as the protection of private data and respect to human rights (Draft AI R&D GUIDELINES for International Discussions, 2017). In the visual text analysis below, you can observe how the values of privacy and autonomy, are connecting (keywords: personal {22, right {25}, respect {10}, autonomous {7}, protection {14}).

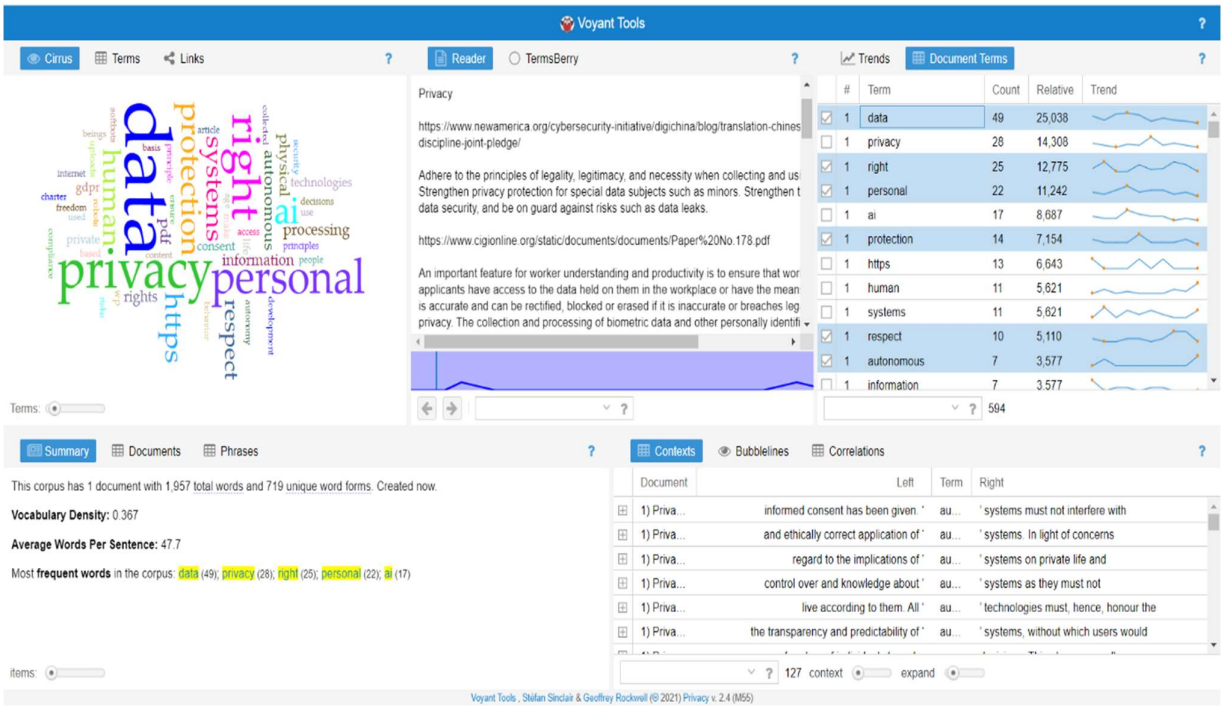


Figure 12, Autonomy/privacy principle, digital text analysis

Furthermore, an example of how an AI system can breach autonomy (free will) and privacy of a person is the WeChat[10] platform in China. WeChat, a messaging app used by millions of people in China, uses automatic analysis to censor text and images within private messaging in real-time. Using optical character recognition, the images are examined for harmful content, including anything about international or domestic politics deemed undesirable by the Chinese Communist Party. It's a self-reinforcing system that's growing with every image sent (O'Neil, 2019).

However, going back to the Amazon Alexa case, one can argue that the company (Amazon), respected the privacy of the user, by enclosing his data, even in the case of murder, confronting with this way a breach of his data privacy. This is the main reason that the normativity here of the Ethical principles of AI, cannot be utilized in any case and precautions on different scenarios as well as legal considerations should be improvised in situations such as this.

Safety and Security

Safety and security are probably the most popular terms when the discussion is emerging about data collection and surveillance, in Artificial Intelligence systems. These principles are

[10] https://www.technologyreview.com/2019/07/15/134178/how-wechat-censors-private-conversations-automatically-in-real-time/. "How WeChat censors private conversations, automatically in real time". July 2019. Accessed November 9th, 2021.

connected to other ethical principles, such as "human rights", "autonomy", accountability, etc.

It is a principle that emerged due to immense and unrestricted use of data, in many cases, without the user's permission. According to the Techno-Anthropological Ethics of Tom Børsen, safety and security can be defined as, "The right to be protected and safeguarded. This value encompasses protection from undesirable events…Safety refers to the right to be safeguarded from unintentional harm, while security refers to the right of protection against intentional harm". (Børsen, 2015:87) Accordingly in the "AI Ethics principles & Guidelines" by Smart Dubai, AI systems should be safe and secure, to protect and serve humanity (Smart Dubai, 2018). An example case of how the safety of humanity can be damaged is the case of the Iranian Scientist killed by a "Machine gun with AI[11] technology" (Kleinman, 2020).

Therefore, here we can witness a different occasion, because safety and security, renowned as "human rights values", are transported to essential ethical principles for the sake of protection from the unrestricted development of Artificial Intelligence. Here, it is an obvious example of how technology society and philosophy are interconnected and influences each other, and how essential is the development of interdisciplinarity.

In conclusion, in the analysis above we witnessed the most popular AI Ethical principles that are appearing in the AI Ethical guidelines database (See Appendix C), their content and how they can be deciphered through the assistance of classical ethical theories and digital methods. In the next chapters, I am going to outline my perspective on how the ethical framework of AI can be developed and what components are missing from the ethics of AI conversation.

[11] https://www.bbc.com/news/world-middle-east-55214359. "Mohsen Fakhrizadeh: 'Machine-gun with AI' used to kill Iran scientist. December 2020. Accessed November 9th, 2021.

6. Discussion

As we witnessed in the previous chapter of the analysis, the predominant values that are being used as the main ethical principles of AI are the aforementioned ones (transparency, fairness, accountability, responsibility, humanity, privacy, safety and beneficence). However, what about the principles that are being the peripheral ones in the conversation, that are missing or overlooked and how we can bridge the gaps in the discussion about a more practical AI Ethical framework? In the following discussion, I will delve into the results of the analysis, the assistance that classical ethical theories and especially notions from Aristotle and Foucault can offer to the conversation and finally, ideas for further research steps.

6.1. Recommendations on AI Ethical principles

After the analysis section and the combining methodology of classical ethical theories and digital text analysis, I am presenting the following table, in which I am examining the content of the aforementioned AI ethical principles and submitting recommendations regarding potential development of these principles.

Table 3, Recommendations on AI Ethical Principles

AI PRINCIPLES	AI PRINCIPLES CONTENT	RECOMMENDATION
Transparency	An AI Pro-ethical condition, that enables the activation of other ethical principles in AI, such as explainability, accountability, privacy, and safety.	An "Ethical Black Box" approach, that will be open to examinations and available in case of bias, thus potential weaknesses can be fixed.
Fairness/Justice	Fairness is being used as an AI ethical principle that is attempting to confront bias & Discrimination (in most of cases, when the AI system has been already utilized).	A Shift to Virtue Ethics approach of Justice, where the developer, will have the phronesis to design the AI system, with precautions (precautionary principle) on potential discrimination and unfairness scenarios
Accountability	The principles of Accountability, connect, transparency with responsibility, seeking an explainable system, thus in case of bias, there should be someone accountable.	An "Ethical Black box" approach, (same with transparency), hence in case of bias, there can be some sort of accountability to the responsible person and to the external factors that led to the decision making.
Responsibility	The principle of Responsibility in AI, seeks to find "who should be responsible for the bias", depending on the consequences of the bias, bringing notions of consequentialist ethics on stage.	A Shift to Virtue Ethics approach of responsibility, that according to Aristotle, there should be an evaluation of the "intentionality" on the decision making. Furthermore, according to Foucault, hierarchical structures should be considered.
Humanity	The "Humanity principle", is concerning human rights, human dignity, and the fact that AI should be secure and safe to humans.	There should be more concrete content on the human rights and legal cautions in case of breaching them and there should be a record to compassion or eudaimonia, on the notion of respecting and helping each other, for the sake of

		humanity. Social and cultural importance is also an essential part of the principle.
Privacy / Autonomy	The privacy/autonomy principle, is based on Normative ethics (Deontological, Consequentialist), depending on when a human's privacy can be violated and can be breached in different scenarios (WeChat example).	Free will (Autonomy), freedom of choice being aware of external power influences as Foucault underlines, can bring autonomy to decision making and protection of privacy. Thus, the AI principle here should be based on the free choice of how someone wants to keep his privacy.
Safety & Security	Safety & Security principle is the one, that focuses on keeping humanity safe from the dangers of AI.	Here I recommend the "Precautionary Principle", which will assist in securing the safe design of AI systems and generate precautions on dangers.
Beneficence	Referred to a high number of codes of conduct, the principle of Beneficence, was described as an evaluation of "more benefit than harm", based on notions from Deontological and Consequentialist ethics.	There is not a universal code of ethics or moral actions, thus it is a necessity to take under consideration, the social and cultural importance and the differences on "What is an ethical action" in different societies or cultures.

Despite the fact that normative guidelines should be supplemented by detailed technical guidance – to the extent that they can be fairly defined – the issue of how to strengthen the precarious situation surrounding the implementation and fulfillment of AI ethics guidelines remains. To answer this question, one must first take a step back and consider ethical theories as a whole. The deontological approach, for instance, is focused on a set of strict laws, responsibilities, or imperatives, while the consequentialist refers to the consequences of every action as the ethical foundation. Character dispositions, moral intuitions, or virtues, on the other hand, are the foundations of the virtue ethics approach. Software developers should follow a collection of basic standards and maxims outlined in ethics guidelines. On the other hand, the virtue ethics approach focuses on "deeper-lying" systems and situation-specific deliberations, as well as personality characteristics and behavioral dispositions among technology developers. Virtue ethics focuses on the citizen rather than on standards of conduct. (Hagendorff, 2020:9).

6.2. Missing elements in the AI codes of conduct

In this chapter, I will clarify the recommendations that I am underlining in table 3 by describing the recommended ethical principles of AI, that I am inclined to believe should be taken into account, during conversations on how the development and the design of AI, could be more responsible and ethical.

Precautionary principle

To begin with, there is a fundamental necessity of setting a framework, before and during the

design of the Artificial Intelligence system, a borderline that will imply limitations and will in a high ratio, identify the consequences, the outcomes of the AI System and will indicate precautions that the developer must take into consideration. Therefore, the precautionary principle is an ideal addition to the AI Ethical guidelines, only referred once in the database, thus in only one code of conduct, there has been a mention of the precautionary principle. In the "table 3", I am clarifying the potential connection of precautionary principles with safety and fairness principles. According to techno-anthropological ethics and Tom Børsen, the precautionary principle is,

> "The principle that states that an action should not be undertaking, if there are reasonable grounds for concern, though no scientific evidence, for it having dangerous effects on the environment, humans, animals or plant health" (Børsen, 2019)

Precautionary principle is more of a guiding principle, whereas the starting purpose of the principle was to encourage decision-makers to take into consideration potential negative effects on the environment, before pursuing these activities (Cameron and Abouchar, 1991). Another definition is coming from Timothy O'Riordan and James Cameron, who are referring Precautionary principle as "…the culturally framed concept that takes its cue from changing social conceptions about the appropriate roles of science, economics, ethics, politics and the law in pro-active environmental protection and management" (O'Riordan and Cameron, 1994:12). Therefore, precautionary principle can be a perfect fit in the Artificial Intelligence systems that are designed, to assist the environment and the sustainable development or confront the global warming, for instance.

Therefore, it is essential for an AI software developer, designer, or engineer to assess and take precautions on the development of an Artificial Intelligence system. Additionally, it is important not solely for negative consequences to humans and bias, such as misinformation, fake news, and discrimination, but furthermore, for the importance of sustainability and respect to nature. Accordingly, it can be combined with another ethical value, deriving from the techno-anthropological ethics, called "Stewardship from the Earth", which is as Tom Børsen referring the responsibility to balance ecosystem resilience and human well-being (Børsen, 2019). Without, taking precautions in AI, it is impossible to secure its ethical development.

Phronesis and Ethical Virtues

A second ethical guideline that can be implemented and combined with AI systems is "Phronesis". The practical rationality of Phronesis that Aristotle defined and I am highlighting in chapter 4.3. of the theoretical framework. In the "table 3", I am connecting phronesis and the ethical virtues with the principles of fairness and responsibility.

According to Aristotle, "the practical rationality "phronesis", is the practical operation of decision that is affecting the ultimate target of "eudaimonia". To achieve "phronesis" one has to possess general knowledge and additionally getting experience from taking decisions considering moral dilemmas, alone or with the advice of the teacher and the laws (Rapp, 2012).

However, one can argue here, that Phronesis cannot be implemented practically in Artificial Intelligence systems, but only as a principle for the person (designer, developer, engineer), that is responsible for the development of the AI technology. Nevertheless, I want to mention here, that if we adjust the theoretical methodology of Aristotle into an AI system, which already possesses a certain amount of general knowledge and train it, in getting experience on decision-making, with the advice of the teacher -who in this occasion in the place of the teacher, will be the software developer, designer or engineer- and reach with that way the stage of practical rationality. In any case, the important factor here is the human and the confrontation of human bias, rather than AI bias. The responsible for the development of technology should be the one, who will be possessing the practical rationality "Phronesis", with a certain level of education and exercise in their life. If we adjust the content of Aristotle's statement, in contemporary theory, a solution would be identified in the right type of "education" and interdisciplinarity. For instance, the data scientist or engineer should have already developed and practiced the way of taking "morally good" decisions into real-time conditions and be prepared to assist the AI system in decision-making and ethical dilemmas. Hence, the practical rationality of "Phronesis", that Aristotle referred to a few thousand years ago, may be a solution for the ethical development of AI systems, today.

Ethical Black Box

"Black Box" is a concept that became popular in airplanes, as a system for the identification of the causes that result in an accident. In a few words, if a flight goes wrong and results in an accident, one can find out, the reasons that the airplane fell or was damaged, by analyzing the content of the "Black box". Accordingly, the term "Black box", revealed in social sciences by Bruno Latour, when he suggested that it is essential to analyze the technology in-depth and not take as, certain the science behind the technology, without examining it. More specifically, Bruno Latour stated that "This is the first decision we have to make: our entry into science and technology will be through the back door of science in the making, not through the more grandiose entrance of ready-made science." (Latour 1987:4). In the "table 3", I am highlighting the importance of an "Ethical black box", for a transparent and accountable ethical development.

Consequently, an "Ethical Black box", can be a reasonably practical ethical framework for Artificial Intelligence systems. UNI Global Union and IEEE have already suggested that AI systems can be equipped with a device called "Ethical Black box", "a device that can record information about said system to ensure its accountability and transparency, but that also includes clear data on the ethical consideration built into the system from the beginning" (UNI Global Union, n.d. as cited in EPRS, 2020:61). Therefore, it can be more concrete in these situations the adjustment of responsibility to certain features and the identification of intentionality. Correspondingly, if we adjust here the Aristotelian notions of intentional or unintentional actions, one could argue that an "Ethical Black Box", can provide some concrete results on the argumentation of which action is intentional and who is responsible. I am recommending the existence of the "Ethical Black Box Principle", as assistance and more concrete framework, to the principles of transparency and accountability.

Social and Cultural Importance

Finally, I am going to state the importance of respecting the social and cultural frameworks. In "table 3", I am underlining the potential connection with the principle of humanity and beneficence. Every culture enhances ethics and moral codes in various ways. For instance, morally "good" actions are being combined a lot of times with religiously "good" actions. Especially in Buddhist societies, it is likely that ethical values should go hand to hand with religion. Here, I want to state that ethics are different from religion, in many ways, at least classical ethics such as Aristotelian and Foucauldian that I am specifically referring to, in this thesis project. However, respecting every culture and identifying how moral codes are interpreted in different societies, is essential for ethics. Furthermore, what I want to underline in this framework of society and culture, is how the feeling of belonging in society and the desire of assisting the society can lead to ethical virtues.

Hence, going back to Foucauldian ethics, I am using his statement, about the importance of social responsibility,

> "The care of the self … implies complex relationships with others insofar as this ethos of freedom is also a way of caring for others... Ethos also implies a relationship with others, insofar as the care of the self enables one to occupy his rightful position in the city, the community, or interpersonal relationships, whether as a magistrate or a friend…"(Foucault, 1997: 287 as cited in Crane et al, 2008:311).

As Foucault states, to reach a stage of caring for yourself, one has to take care of the others at the same time and reach a stage of assisting the whole community. Consequently, Aristotle emphasized the importance of ethical actions that are assisting the whole society by stating that "the quest of "eudaimonia" is opposing the selfishness at the expense of others and secondly, recognizes that the prosperity and welfare of an individual are connected with the prosperity and welfare of the society" (Rapp, 2012:54).

In conclusion, prosperity and welfare of society -and not only just benefit for the individual- is essential for the development of ethics. Accordingly, to reach the ethical virtue not only for the individual but for the whole society, one should have already recognized and respected what each society and culture is conceiving as ethical virtue. Hence, for the concept of Artificial Intelligence, the AI systems, have to be adjusted in different societies and cultures and not use universal ethical codes, or global AI Ethical principles..

6.3. Power Structures on Decision Making

In this chapter of the discussion, I want to highlight the importance of power structures and domination in the development of the AI Ethical codes of conduct, underlining arguments from the Foucauldian Ethics, that I am stating in chapter 4.2. of the theoretical framework. In the following visualization, you can witness a global map of the countries that have introduced codes of conduct on Ethical AI development.

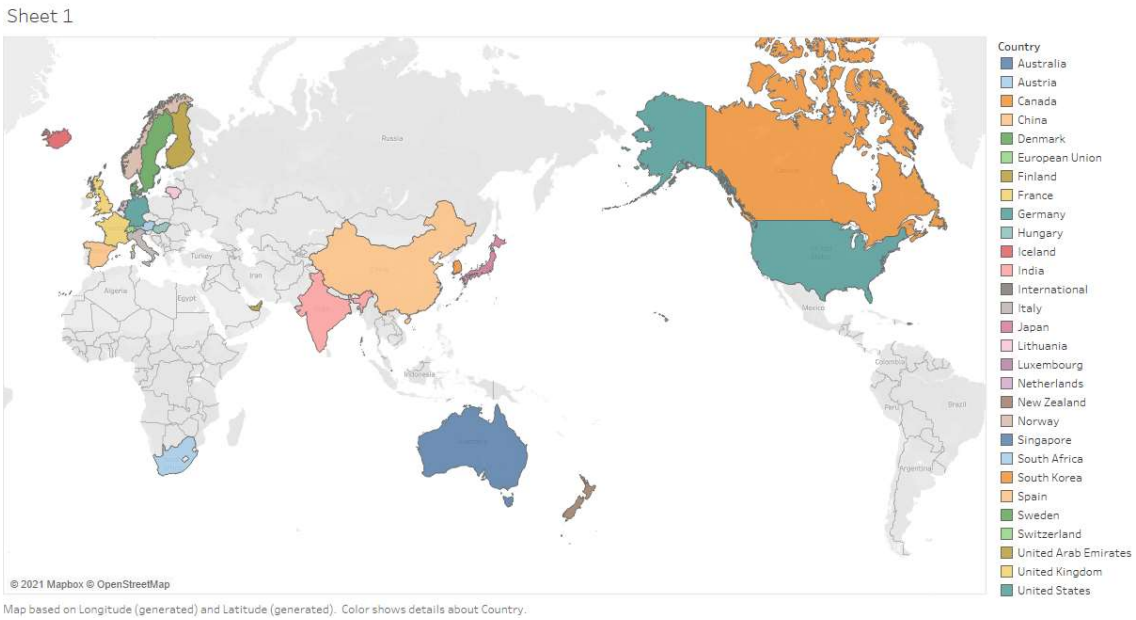


Figure 13, Map of countries that developed AI Ethical codes of conduct, made by Tableau

In the data visualization above, one can witness the inequality of publications in western and non-western societies. This exclusion of the non-western societies arises questions on the universality of the AI Ethical principles. A universal code of ethics is impossible due to the cultural and societal differences that must be considered; thus, this attempt at universality is problematic. Accordingly, Foucault referred to ethics as a practical concept, that is not following religious or universal norms, instead, ethics are used to denote the individual agency possibilities (Crane et al, 2008). Undoubtedly, there are a few dominating actors (US, UK, Germany, France), who are "ruling" the development of the Ethical AI codes of conduct, having almost a "monopoly" in the publications, excluding non-western societies from the dialogue. A unique, perfect communication between different societies was always difficult. However, Foucault as mentioned by Krane, defines that these power games can be played with a minimum amount of domination, only if there is the opportunity "to give one's self the rules of law, the techniques of management, and also the ethics, the ethos, the practice of self" (Foucault, 1994: 18 as cited in Krane et al, 2008:304). Hence, a hierarchical power structure, will continue to exist, and the other actors (non-western societies), should not be unaware of this power game, but instead raise their voice and attempt to deliver their own perspectives to the table, resisting the hierarchical domination.

Additionally, another "hidden power game", regarding the production of AI codes of conduct, exists in the imbalance of publishments between academia and governmental papers or the private sector. Checking figure 5, in chapter 5.1., you can witness that only 18 from 162 codes of conduct, have been published solely from Academia, while 44 are governmental papers and 38 codes have been developed from private sectors and civil society organizations, correspondingly. As governments and private sectors, seem to have the power and political domination on the development of these codes, questions about the interests of the various stakeholders (big-tech companies, political parties) are rising. Ethics should not represent the interest of a few members of society, but the whole society. Accordingly, for Habermas ethics are residing in "the institutionalization of... procedures and conditions of

communication" in democratic decision-making processes, which means that the public use of corporate power, should be under democratic control (Scherer & Palazzo, 2007 as cited in Crane et al, 2008:308). In conclusion, the Foucauldian Ethics, are highlighting the attention on the power domination in the development of an AI framework, and how ethics can be used as a confrontation of the individual agency.

6.4. A shift to education and next steps

In the last chapter of discussion, I want to highlight the essential part of ethics in education and the imminent shift to Virtue Ethics and their implications in AI, which can only be a successful concept, through a strong focus on ethics education, deriving from Aristotelian notions of ethics. Ethics lectures and conversations, both at educational institutions and as postgraduate continuing education, are critical in making participants aware of the challenges, even if they do not have the solution. Ethics should not be just a technical code or an add-on enforced into technical solutions, it should be embedded into the culture of technical systems and the education of the one who is responsible for the development and the design of AI systems.

Additionally, a second statement from classical ethics, regarding phronesis is that

For an engineer, programmer, or other designer involved in the AI movement as it gains traction in the future, making the right decisions is critical. As I referred on chapter 4.3.“… applied ethics and philosophy is not something that can be randomly taught to someone, instead, it is an ability that one can possess with experience and practical dilemmas (Rapp, 2012:20).” However, AI tasks such as predicting all potential situations and determining what decisions should be taken in each will be undeniably, difficult. Ethics courses and discussions, both in educational institutions and as postgraduate continuing education, are critical in making participants aware of the problems, even if they do not know the answers.

To mitigate a variety of issues and identify goals, it is important to explain the vision and scope of this part of their education. To begin with, ethics cannot be reduced to basic values or doctrines that can be placed on students in order to turn them into "good people.". Moreover, it does not include teaching conformity lessons that solely focus on adhering to all applicable laws and regulations as outlined in company policies; we already expect tech experts to follow the rule. The goal of teaching ethics is to provide the intellectual resources that potential architects of a digital society would need to be able to recognize and deal with moral issues that they confront (Villani, 2018:123).

The reason I am emphasizing the education in ethics, or ethical principles or ethical virtues, is specifically because throughout the centuries, starting with Aristotelian and Virtue Ethics, there was always the notion, that “Ethos”, the repetitive practice of moral action, can lead to the acquisition of the ethical and intellectual virtues. Furthermore, guidance from the educator (pedagogist (παιδαγωγός in Greek for Aristotle)), was a requirement on learning how to maintain ethical decision making. “To achieve “phronesis” one has to possess general knowledge and additionally getting experience from taking decisions considering moral dilemmas, with the advice of the teacher and the laws (Rapp, 2012:20)”. Accordingly, the bias most of the time is not a responsibility of the Artificial Intelligence system/software, but

instead, it is likely to be a result of human's fault. Human bias is what we must target and education in ethics is arguably a solution that can assist the development of ethical virtues to the software developers, engineers, etc. There would therefore seem to be a definite need for an interdisciplinary education that combines different sectors of science. Aristotelian (virtue) ethics showed us the way, long time ago. The importance of interdisciplinarity in education and the combination of technological studies with social science studies are absolute.

Franz Boas one of the founders of Anthropology once wrote, "we have simple industries and complex organization" and "diverse industries and simple organization" when comparing the structures of societies with simple tools in comparison to those with seemingly complex technologies (Royer, 2020:17). This statement represents a central idea in the field, which is to turn our assumptions about how our societies work their structure, class structures, and general organization (Royer, 2020). Artificial intelligence, according to anthropologists like Hagerty and Rubinov, is a socio-technical construct, which means that "the technological aspects of AI are intrinsically and closely related to its social aspects" (Royer, 2020:18). With this statement, I am concluding the discussion part, with the intention to highlight the importance of social sciences, interdisciplinary studies such as Techno-anthropology and education on ethics, as essential factors for the building of a responsible and ethical AI framework.

7. Conclusion

The aim of this project is not a critique of the development of Artificial Intelligence, but rather a research, on the ethical framework of Artificial Intelligence, with a twofold target. Attempting the development of AI Ethical principles by firstly, understanding the content of the contemporary ethical codes of conducts of Artificial Intelligence and secondly, analyzing the AI ethical principles that have been developed from various institutions by exploring classical ethical theories and a digital text analysis approach.

Inspired by the interdisciplinary field of Techno-Anthropology I developed a mixed-methods approach to answer these inquiries. Firstly, digital methods, to collect AI Ethical codes of conduct to generate a database, that I could explore and identify the "AI Ethical" principles and secondly, digital text analysis to isolate the keywords on the certain definition provided by various institutions on the AI ethical principles and analyze them, under the assistance of classical ethical theories.

As discussed in the introduction part, in the last few years, the discussion for AI Ethical guidelines or principles has been increased (Jobin et al, 2019). Transparency, autonomy, security and safety, justice, accountability, Human dignity, are some of the top principles in the AI Ethical framework, discussions. However, these codes seem to recycle specific theories and principles, avoiding the practicalities of technical codes. Additionally, they seem to hunt a universal AI Ethical framework, omitting the cultural and social aspects of humanity. A possible explanation for these results may be the lack of concrete ethical theories behind them and the power-political structures and hierarchical dominations that are forcing the development of certain ethical frameworks in AI systems.

This thesis paper is proposing a shift from the Normative, Deontological ethics and abstract AI Ethical principles to Aristotelian Virtue Ethics with the addition of a social and cultural framework and design processes such as the precautionary principle and the ethical black box. This suggestion must be approached with some caution because the main recommendation here, is that this change of direction, needs to start from an educational level. The interdisciplinarity in education and the connection of engineers with philosophers, data scientists with anthropologists can be the starting point for this shift, whereas the essential part is that the developers of AI systems, should possess a general knowledge of ethical theories and obtain experience in the confrontation of bias and moral dilemmas.

Furthermore, the second main point of this thesis is that the Ethical frameworks and guidelines for AI, should not solely be implemented by Big-Tech companies or exclusively from "western societies". A reasonable approach to tackle this issue could be the establishment of a constant dialogue, which involves different hierarchical groups, that will develop the awareness of minor actors in the political-power game about the "underground" power structures. At the same time, the conversation and inclusion of excluded-oppressed groups, in the conversation can offer innovative perspectives on the implementation of an equally Ethical framework in AI.

I am inclined to believe that this thesis project under the discipline of Techno-Anthropology, with the combination of Aristotelian and Foucauldian ethical theories, can contribute to the

issue of AI Ethics implementation, which is as pressing, unpredictable, and divisive as it has always been. The development and the constant reproduction of the knowledge with a unique methodology such as the mix-methods interdisciplinary approach that Techno-Anthropology is utilizing can offer innovative insights.

References

Athanasopoulos, Panagiotis. “Ethika Nikomacheia, text, translation, comments, inquiries, cases” (Ancient Greek: “Ἠθικὰ Νικομάχεια, κείμενο, μετάφραση, σχόλια, ερωτήσεις θέματα εξετάσεων”) (2013 – 2014)

Bartneck, Christoph, Christoph Lütge, Alan Wagner, and Sean Welsh. An Introduction to Ethics in Robotics and AI. SpringerBriefs in Ethics. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-51110-4. (2021)

Boddington, Paula. Towards a Code of Ethics for Artificial Intelligence. Artificial Intelligence: Foundations, Theory, and Algorithms. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-60648-4. (2017).

Børsen, Tom, and Søren Nors Nielsen. 2017. "Applying an Ethical Judgment Model to the Case of DDT." HYLE–International Journal for Philosophy of Chemistry 23 (2017): 5-27.

Børsen, Tom.“Preparing for Ethical Judgment in Techno-Anthropology,Techno-Science and Engineering.” In Proceedings from the 41th SEFI Conference , 16-20 September 2013, Leuven, Belgium European Society for Engineering Education. (2013)

Botin Lars. Pernille Berlersen and Christian Nøhr. “Techno-Anthropology in Health Informatics. Methodologies for improving human-technology relation.” (2015)

Buruk, Banu, Perihan Elif Ekmekci, and Berna Arda. "A critical perspective on guidelines for responsible and trustworthy artificial intelligence." Medicine, Health Care and Philosophy 23, no. 3 (2020): 387-399.

Cameron, James, and Juli Abouchar. "The precautionary principle: a fundamental principle of law and policy for the protection of the global environment." BC Int'l & Comp. L. Rev. 14 (1991): 1.

Chatila, Raja, and John C. Havens. “The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.” In Robotics and Well-Being, edited by Maria Isabel Aldinhas Ferreira, João Silva Sequeira, Gurvinder Singh Virk, Mohammad Osman Tokhi, and Endre E. Kadar, 95:11–16. Intelligent Systems, Control and Automation: Science and Engineering. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_2. (2019).

Christensen Mads, Amit Maor, and Georgios Natsios. “Moral Deskilling in Urban Planning: A Techno-Anthropological Commentary on Ethics in a Digital Development” (2019)

Christof Rapp. “Introduction to Aristotle” (translation Hlias Tsirigkakis) Athens: Ochto (8) Editions (Greek “Εισαγωγή στον Αριστοτέλη”. Christof Rapp. Μετάφραση Ηλίας Τσιριγκάκης. (2012)

Cingref and Syntec Numerique. “DIGITAL ETHICS: A GUIDE FOR PROFESSIONALS OF THE DIGITAL AGE” (2018)

Crane, Andrew, David Knights, and Ken Starkey. “The Conditions of Our Freedom: Foucault, Organization, and Ethics.” Business Ethics Quarterly 18 (3): 299–320. https://doi.org/10.5840/beq200818324. (2008)

Crockett Molly. “The trolley problem: would you kill one person to save many others?” (2016)

CSSF.” Artificial Intelligence: OPPORTUNITIES, RISKS AND RECOMMENDATIONS FOR THE FINANCIAL SECTOR” (2018)

D’Acquisto, Giuseppe. "On conflicts between ethical and logical principles in artificial intelligence." AI & SOCIETY 35, no. 4 (2020): 895-900.

Dubai, Smart. "AI ethics principles & guidelines." Smart Dubai Office (2019).

European Group on Ethics in Science and New Technologies. "Statement on artificial Intelligence, robotics and ‘autonomous’ systems." Retrieved September 18 (2018)

European Parliament. Directorate General for Parliamentary Research Services. The Ethics of Artificial Intelligence: Issues and Initiatives. LU: Publications Office. https://data.europa.eu/doi/10.2861/6644. (2020)

Hagendorff, Thilo. “The Ethics of AI Ethics: An Evaluation of Guidelines.” Minds and Machines 30 (1): 99–120. https://doi.org/10.1007/s11023-020-09517-8. (2020)

Haselager, Pim, and Giulio Mecacci. "Superethics instead of superintelligence: know thyself, and apply science accordingly." AJOB neuroscience 11, no. 2 (2020): 113-119

Hauer, Tomas. "Machine Ethics, Allostery and Philosophical Anti-Dualism: Will AI Ever Make Ethically Autonomous Decisions?." Society 57, no. 4 (2020): 425-433.

IBE. “Business Ethics and Artificial Intelligence”. Business Ethics Briefing. Issue 58. (2018)

Jobin, Anna, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines." Nature Machine Intelligence 1, no. 9 (2019): 389-399.

Johnson, Arthur T. “Ethics in the Era of Artificial Intelligence.” IEEE Pulse 11 (3): 44–47. https://doi.org/10.1109/MPULS.2020.2993667. (2020)

Larsson, Stefan. “On the Governance of Artificial Intelligence through Ethics Guidelines.” Asian Journal of Law and Society 7 (3): 437–51. https://doi.org/10.1017/als.2020.19. (2020)

Latour, Bruno. Science in Action: How to Follow Scientists and Engineers through Society. Cambridge, Mass:Harvard University Press. (1987)

Leslie, David. "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector." Available at SSRN 3403301 (2019).

Luccioni, Alexandra, and Yoshua Bengio. "On the morality of artificial intelligence [Commentary]." IEEE Technology and Society Magazine 39, no. 1 (2020): 16-25.

Markopoulos Ioannis. “Science - Technology and philosophical reflection: historical overviews and epistemological and evaluative correlations” -Greek Translation-

“Μαρκοπουλος, Ιωαννης. Επιστημη Τεχνολογια και φιλοσοφικος στοχασμος: ιστορικες επισκοπησεις και γνωσιολογικες και αξιολογικες συσχετισεις. University Studio Press, 2018.

McLaren, Bruce M. "Extensionally defining principles and cases in ethics: An AI model." Artificial Intelligence 150, no. 1-2 (2003): 145-181.

O'Riordan, Timothy, and James Cameron. Interpreting the precautionary principle. Routledge, 2013.

Royer, Alexandrine. "The Short Anthropological Guide to the Study of Ethical AI." arXiv preprintarXiv:2010.03362 (2020).

Stilgoe, Jack, Richard Owen, and Phil Macnaghten. "Developing a framework for responsible innovation."In The Ethics of Nanotechnology, Geoengineering and Clean Energy, pp. 347-359. Routledge, 2020.

United Nations and Institute of Macau. N.d. “A typological Framework for data marginalization”.

Villani, Cedric. For a Meaningful Artificial Intelligence: “Towards a French and European strategy” (2018)

Wang, Amy. “Can Amazon Echo help solve a murder? Police will soon find out.” The Washington Post. (2019)

Zins, Chaim. "Success, a structured search strategy: Rationale, principles, and implications." Journal of the American Society for Information Science 51, no. 13 (2000): 1232-1247.

Internet Specific sources

Algorithmic Watch – “AI Ethics Guidelines Global Inventory”: https://inventory.algorithmwatch.org/ . April 2020. Accessed April 20, 2021

BBC. “Mohsen Fakhrizadeh: 'Machine-gun with AI' used to kill Iran scientist. https://www.bbc.com/news/world-middle-east-55214359 . December 2020. Accessed November 9, 2021. (2020)

Børsen, Tom. Presentation about Ethical “Techno Anthropological Problems and Theories”: https://www.moodle.aau.dk/course/view.php?id=30584#section-14 . November 2019. Accessed November 14, 2021. (2019)

Børsen, Tom. Presentation about RRI “Techno Anthropological Problems and Theories”: https://www.moodle.aau.dk/course/view.php?id=30584#section-14 . November 2019. Accessed November 14, 2021. (2019)

Britannica “logos”. https://www.britannica.com/topic/logos . N.d. Accessed June 3, 2021

Byler, Darren. “China’s hi-tech war on its muslim minority“ https://www.theguardian.com/news/2019/apr/11/china-hi-tech-war-on-muslim-minority-

xinjiang-uighurs-surveillance-face-recognition . April 2019. Accessed, November 8, 2021. (2019)

Cook, James. "Amazon scraps sexist AI recruiting tool showed bias against women" https://www.telegraph.co.uk/technology/2018/10/10/amazon-scraps-sexist-ai-recruiting-tool-showed-bias-against/. October 2018. Accessed, November 7, 2021. (2018)

Europa "Tax Edu": https://europa.eu/taxedu/young_el . June 3, 2021. Accessed June 3, 2021

Github - "Awful AI": https://github.com/daviddao/awful-ai . December 30, 2020. Accessed May 1, 2021.

Hunt, Elle. "Tay microsofts AI chatbot gets a crash course in racism from twitter". https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter. March 2016. Accessed, November 7, 2021. (2016)

Instant data scraper. https://chrome.google.com/webstore/detail/instant-data-scraper/ofaokhiedipichpaobibbnahnkdoiiah . (Data scraping)

ITU. "Artificial Intelligence for good". https://www.itu.int/en/mediacentre/backgrounders/Pages/artificial-intelligence-for-good.aspx. June 2021. Accessed, November 4, 2021. (2021)

Levin, Sam. "New Artificial Intelligence app can tell whether you are gay or straight from a photograph". https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph. September 2017. Accessed, November 7, 2021. (2017)

Marr, Bernard. "How artificial Intelligence, IOT and big data can save the bees". https://www.forbes.com/sites/bernardmarr/2020/04/22/how-artificial-intelligence-iot-and-big-data-can-save-the-bees/?sh=17784a3d1d9e. April 2020. Accessed, November 4, 2021.

O'Neil, Patrick Howell. "How Wechat censorts private conversation automatically in real time". https://www.technologyreview.com/2019/07/15/134178/how-wechat-censors-private-conversations-automatically-in-real-time/. July 2019. Accessed November 9, 2021. (2019)

Pownall, Augusta. "Huawei facing emotion app uses sound to allow the visually impaired to see emotions. https://www.dezeen.com/2019/01/02/huawei-app-blind-facing-emotions/. January 2019. Accessed, November 4, 2021. (2019)

Rijnemann, Van Mark. 2020. "Why We Need Ethical AI: 5 Initiatives to Ensure Ethics in AI" https://vanrijmenam.nl/why-we-need-ethical-ai-5-initiatives-ensure-ethics-ai . January 2020. Accessed June 1, 2021. (Used for the logo)

Tableau Software. https://www.tableau.com/ . (Data visualizations)

Telia company. "AI Ethics". https://www.teliacompany.com/en/about-the-company/public-policy/ai-ethics/. N.d. Accessed, November 7, 2021.

The trolley problem: https://www.theguardian.com/science/head-quarters/2016/dec/12/the-trolley-problem-would-you-kill-oneperson-to-save-many-others . December 2016. Accessed June 3, 2021

The World Bee Project. https://worldbeeproject.org/. n.d. Accessed November 4, 2021.

Vincent, James. "Service that uses AI to identify gender based on names looks incredibly biased". https://www.theverge.com/2020/7/29/21346310/ai-service-gender-verification-identification-genderify. June 2020. Accessed, November 8, 2021. (2020)

Vincent, James. "What a maching learning tool that turns Obama white can (and can't) tell us about AI bias".https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias. June 2020. Accessed, November 7, 2021. (2020)

Voyant tools. https://voyant-tools.org/ . (Digital Text analysis)

Woiceshyn, Jaana. "Why Businesspeople need moral principles". https://www.nassauinstitute.org/why-businesspeople-need-moral-principles/. June 2018. Accessed, November 14, 2021. (2018) (Used for the logo).

Acknowledgements

I would like to thank my supervisor, Professor Tom Børsen, whose expertise was valuable in completing my thesis report. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.