Exploring How Developers Could Include the European Commission's Ethics Guidelines to Strive Toward Trustworthy AI

Esmée Stouten
Graduation Thesis, April 2019
Media Technology MSc program, Leiden University

Exploring How Developers Could Include the European Commission's Ethics Guidelines to Strive Toward Trustworthy AI

Esmée Stouten

Graduation Thesis, April 2019

Supervised by Maarten Lamers[1], Maaike Harbers[2] and Peter Paul Tonen[3]

Media Technology MSc program, Leiden University

esmeestouten@gmail.com

[1] Leiden University, Leiden Institute of Advanced Computer Science, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands

[2] Rotterdam University of Applied Sciences, Kenniscentrum Creating 010, Wijnhaven 103, 3011 WN Rotterdam, The Netherlands

[3] Capgemini Netherlands, Applied Innovation Exchange, Reykjavikplein 1, 3543 KA Utrecht, The Netherlands

Universiteit Leiden
Leiden Institute of Advanced Computer Science

Capgemini
APPLIED INNOVATION EXCHANGE

Abstract

Artificial intelligence (AI) is neither a revolution nor hype. As the potential benefits of AI are increasing, the usage of AI-enabled technologies also engenders certain threats and risks that require a careful approach. Based on a first draft of ethics guidelines concerning trustworthy AI from the High-Level Expert Group on AI, this explorative study attempts to provide critical insight into the way companies could offer their developers guidance regarding the concrete operationalization of trustworthy AI. To this end, during a time frame of nine months, two qualitative studies were conducted. First, semi-structured interviews were held with 16 stakeholders to map current problems and possible solutions within the AI fields of different companies. Subsequently, developers for the IT-consulting corporation Capgemini were asked to utilize the assessment list designed by the European Commission to gain an understanding of how their current working method corresponds with the key components of the original list. The results demonstrate that the subjectivity in defining ethical concepts and the unique view that every person possesses on an ethical dilemma make it difficult to compose a predetermined set of rules for making informed choices that adhere to certain principles. Also, the fuzzy front-end of innovation processes as well as the dynamics within AI-enabled technologies ensure that an assessment cannot be truly comprehensive. Additionally, composing rules when a project's direction is yet unknown ensures that there can be no stability in this regulation. Besides these content-related difficulties, there appeared to be tension between gaining freedom and taking responsibility; developers usually do not look beyond optimizing technical possibilities. In response to these overarching difficulties, recommendations to company-specific guidelines have been provided; among others, avoid ambiguity by giving meaning to subjective terms, utilize examples to illustrate relevance when considering a specific aspect, and add a weighted-scale checklist to gain insight into the extent to which a system is compliant with ethics guidelines. As little research has been conducted into drafting ethical guidelines regarding AI for companies, it cannot be claimed that different forms of ethical testing are not at least as effective as composing ethical guidelines. It is therefore important that more research is conducted toward applying a system of ethical governance in AI and robotics.

Keywords: Artificial intelligence, European Commission, ethics, guidelines, trustworthy, developers, responsibility, subjectivity.

Table of Contents

1 Introduction

Artificial intelligence (AI) has been around for decades. As stakeholders influence technological innovations by deciding whether a development satisfies their needs and goals, they have obtained a prominent role in transforming digital society (Davenport & Katyal, 2018; Smuha, 2018). An overall potential of those AI-enabled technologies is that tasks that previously could only be performed by humans can be supplemented, substituted, and amplified, which increases the prosperity and welfare of society (Makridakis, 2017; Scherer, 2015; Smuha, 2018). Optical character recognition, for instance, can convert written or printed words into data; autonomous machines can execute complex financial transactions; and algorithms are able to process significant quantities of documents within a few seconds (Scherer, 2015). Artificial intelligence has, therefore, "the capability to generate benefits for individuals and society" (Smuha, 2018, p. i).

While the usage of AI-enabled technologies could play a key role in a company's effort to cut costs and increase its economic competitiveness, these same technologies are also an essential element for companies across different industries to drive revenue and increase profits (Makridakis, 2017). Scherer (2015) mentions that "[it] gains [a] strong foothold in [industries and firms] and becomes more enmeshed in our day-to-day lives, and that trend seems likely to continue for the foreseeable future" (p. 354). It seems that the competitive position of a company depends largely on the extent to which it is able to generate useful insights. The usage of robotics and AI enables companies to solve business problems, eliminate routine tasks, and promote efficiency. KLM Royal Dutch Airlines, for instance, is contacted over 130,000 times per week on Twitter, Messenger, and WhatsApp. By utilizing AI and supporting service with technology, they can handle more questions in a shorter period of time, and their conversations with customers become more efficient, relevant, and personal than competing companies (KLM, 2017). Another example of an organization that utilizes AI to strengthen the position of their business is the Dutch police department Q. Q utilizes AI and machine learning to find possible DNA evidence in cold cases but attempts to adapt the technology to recognize other forensic evidence and hopes that the technology will eventually identify non-forensic evidence as well (Verhagen, 2018). These examples illustrate that by deploying the appropriate AI applications and by knowing which technologies provide maximum benefits, programmers can enable AI to be beneficial in different domains, such as engaging customers and transforming products or services (Makridakis, 2017).

Although AI offers many potential benefits, it also engenders certain threats and risks that require a careful approach. Examples from the recent past illustrate that the introduction of IT applications is not insensitive to legal and ethical debates (Smuha, 2018). Amazon's AI recruiting tool, for instance, was trained to rate candidates for software development jobs and other technical posts by observing patterns in their 10-year-old database that was filled with old resumés. As a consequence of the male dominance across the technology industry, the AI became unfair through the biased training dataset it used. The recruiting tool was not sorting candidates in a gender-neutral manner, and resumés that included any references to women were penalized, such as the Women's College of [university name] (Dastin, 2018). In addition, the Facebook-Cambridge Analytica data scandal demonstrated the possible risks of technology on citizens' privacy. Isaak and Hanna (2018) mention that Cambridge Analytica, a company that combined data analysis with strategic communication, identified voters who might be enticed to vote for their client or discouraged from voting for their client's opponent: "They developed the ability to 'micro-target' individual consumers or voters with messages most likely to influence their behavior" (p. 57). These examples exhibit that AI must be developed in a careful manner.

There are two main groups that can take responsibility in realizing AI: regulating authorities, such as the government, and relevant stakeholders who are developing, deploying, or utilizing AI, which encompasses companies, public services, researchers, and individuals. Scherer (2015) mentions that it is a challenge for regulatory authorities to make precise legislative definitions of AI, because "any legal definition for the purposes of liability or regulation likely would be over- or under-inclusive [. . . and] courts have always needed to adjust the rules for proximate causation as technology has changed and developed" (p. 373). This means that companies that have embedded ethics at the hearts of their organizations are currently facing an important challenge in applying AI responsibly: where do they and their employees draw the line between what they want to develop, should develop, and could develop according to the law?

A system of ethical governance in AI and robotics could be applied in different ways. Luxton (2014) and Winfield and Jirotka (2018) argue, for example, that organizations must be transparent regarding ethical governance so that it becomes part of the organization's DNA. Torresen (2018) indicates that developers need to be aware of possible ethical challenges and that they should address ethical issues in the design of their AI systems, including avoiding misuse and respecting human autonomy. By ensuring professional guidelines and an ethical code of conduct that addresses ethical risks, an organization's expectations become clear: "Ethics and responsible innovation, like quality, is not something that can be implemented as an add-on; simply appointing an ethics manager [. . .] is not enough" (Winfield & Jirotka, 2018, p. 10). Besides this governmental regime that desires to make those involved consider ethical issues, others advocate public regulators who participate in and interact within the IT field by "[gathering] information and knowledge about the industries [. . .] and [classifying] various risk categories" (Guithot, Matthew &

Suzor, 2017, p. 2). These varied solutions illustrate that it is difficult to define how these institutional ethical policies must be translatable into concrete practices (Boddington, p. 34).

The High-Level Expert Group on AI from the European Commission contemplated the challenge of applying AI responsibly and provided a first draft of ethics guidelines concerning AI to "offer [stakeholders] guidance on the concrete implementation and operationalization" of core values and principles for trustworthy AI (Smuha, 2018, p. i). As advice to increase the striving toward trustworthy AI, the European Commission suggests several methods that companies can employ. One of them is to adapt their codes of conduct, or charters of corporate responsibility, in such a way that: "an AI system can [. . .] document its intentions, as well as underwrite them with standards of certain desirable values such as fundamental rights, transparency and the avoidance of harm" (Smuha, 2018, p. 22). By emphasizing that an assessment list cannot be exhaustive, the aforementioned problem of a lack of comprehensiveness for technological regulation is confirmed. Any assessment list must therefore be adapted "to the specific case in which the system is used [. . . which is] a continuous process of identifying requirements, evaluating solutions and ensuring improved outcomes throughout the entire lifecycle of the AI system" (Smuha, 2018, p. iii).

Therefore, the main question this research project aims to answer is this: How could developers include the European Commission's ethical guidelines to increase the striving toward trustworthy AI? In exploring this question, the following objectives are met:

1. Gain an understanding of how developers currently ensure that their AI development maximizes benefits while minimizing risks.
2. Discover the extent to which developers currently consider the European guidelines and explore areas that need attention to ensure that developers are aware of and trained in trustworthy AI.

In the remainder of this paper, the answer to the main question is provided utilizing the following structure. The second section provides the theoretical background to the research, briefly outlining the obstacles in the research. Also, a short summary of the European Commission's document is provided. In Section 3 the method and procedures for this study are explained; both semi-structured interviews and the concept, functionalities, and design of the ethics guidelines gained from consultation with the target group are provided. Results of stakeholders' reflections on current problems and solutions within the AI field and developers' interpretations of the assessment list are revealed in Section 4. The insights gained from both studies have been compared to locate obstructions that can be resolved at both the organizational and the content levels. Recommendations based on these overarching difficulties are supplied in Section 5. In Section 6, fundamental and resource-driven limitations of the study are considered, and recommendations for further research are presented. Section 7 concludes the research. To guarantee the privacy of all stakeholders, the appendices have been added separately.

2 Brief Background of Relevant Factors

In this section, a short description of the European Commission's guidelines for stakeholders is furnished, and their working definitions are described. Subsequently, literature from different fields is discussed to illustrate how this paper delineates the term *AI*. As ethical values play a major role in this study, a fresh look must be taken regarding companies' propagation of their ideologies toward employees. Last, key findings are submitted.

2.1 The European Commission's Requirements for Trustworthy Artificial Intelligence

The document of the European Commission regarding AI consists of three layers. These guidelines are summarized here so that the research foundations for this study are clear. As this section does not utilize any further sources or references, for all quotations and specific terminology (Smuha, 2018) refer to the European Commission's guidelines. Hence, no further references are made.

In the first chapter of their document, the fundamental rights on which European principles and values are built are tailored to cover the AI field. They consist of five families of rights: respect for human dignity; freedom of the individual; respect for democracy, justice, and the rule of law; equality, non-discrimination, and solidarity, including the rights of persons belonging to minorities; and citizen rights (p. 7). Five principles arise from these fundamental values. To begin with, AI should be developed to meet the world's great challenges and to bring more goodness into the world (e.g., providing solutions for climate change by optimizing energy efficiency). Second, AI development should be designed to avoid harm in any case toward humans, the environment, and animals. Third, AI-enabled technologies must be subordinate to humans; for example, the right to self-determination for humans is obliged. Fourth, every individual must have equal opportunities in terms of access to services and technology by avoiding bias, stigmatization, and discrimination against minorities. Fifth, the operationalization of any AI system, the associated business model, and the intention of its developers must be sufficiently transparent to be auditable, comprehensible, and intelligible by humans (p. 8-10).

The second chapter proposes 10 requirements that AI must meet to comply with these principles, which are based on the foregoing fundamental rights. These requirements are displayed in Table 1. Moreover, technical and non-technical solutions are furnished to ensure that AI is built in a human-centric manner and that the developments are built upon fundamental rights, principles, and values (p. 18-22).

The third chapter proposes the usage of an assessment list to operationalize these requirements throughout every step of the design and development of an AI-enabled technology. The draft version that the High-Level Expert Group on AI offered serves as the starting point for this study.

Table 1. The requirements for trustworthy AI in alphabetical order. A detailed description of the requirements is given on p. 14-18.

Accountability (ACC)	Respect for Human Autonomy (RFHA)
Data Governance (DG)	Respect for Privacy (RP)
Design for All (DFA)	Robustness (R)
Governance of AI Autonomy (GAA)	Safety (S)
Non-Discrimination (ND)	Transparency (T)

Finally, the High-Level Expert Group on AI has added a glossary in which they discuss various definitions, two of which are presented below and are built upon later in this study.

> *Artificial intelligence* are systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best actions to take, according to predefined parameters, to achieve the given goal (p. iv).
>
> *Trustworthy AI* should respect fundamental rights, applicable regulation and core principles and values, ensuring an ethical purpose. [Moreover], it should be technically robust and reliable since, even with good intentions, a lack of technological mastery can cause unintentional harm (p. iv).

2.2 An Unclear Vocabulary: Defining Artificial Intelligence

Although the European Commission has provided a clear definition for AI, the literature demonstrates that there is little consensus on its meaning. Since there is no definition of intelligence that does not tend to be linked to human characteristics, it has become difficult to recognize intelligent agents as possessing intelligence (Scherer, 2015). In fact, AI definitions are often tied to the ability to perform specific intellectual tasks (Scherer, 2015; LaChat, 1986). Gurkaynak, Yilmaz, and Haksever (2016), LaChat (1986), and Scherer (2015) argue, therefore, for an AI term that is not linked to human characteristics because machines can already outperform humans in specific tasks, such as information retrieval and performing repeated actions.

Additionally, finding something intelligent depends on who or what performs a specific intellectual task (Harbers, 2018). For example, there appears to be a distinction between intelligent behavior and personally intelligent behavior: "A machine can learn to run a maze as well as a rat and at the level of rat intelligence could be said to be fairly 'smart'" (LaChat, 1986, p. 72). Considering that an agent's unique skill set is only a small segment of the capacity of human intelligence because it is highly specialized, one may wonder if intelligence should be tied to the ability to perform tasks at all.

Moreover, milestones in the AI field may seem ordinary once technological advances enable agents to perform those activities that were previously seen as groundbreaking (LaChat, 1986; Scherer, 2015). Scherer (2015) states that "[milestones have] not been to proclaim that the machine that achieved it possesses intelligence, but rather to interpret the accomplishment of the milestone as evidence that the trait in question is not actually of intelligence" (p. 361). The conceptual ambiguity of the term *intelligence* and the continuous shift for finding AI-enabled technologies intelligent makes it difficult to devise a timeless definition for AI.

Nowadays, various terms of AI are employed to describe an agent's intelligence in comparison with human intelligence. Turing stated in his paper *Computing Machinery and Intelligence* that "if [a] machine successfully pretends to be human to a knowledgeable observer [. . .] should consider it intelligent" (Turing, 1950). According to McCarthy (1998), intelligence is "the computational part of the ability to achieve goals in the world," and AI is "the science and engineering of making intelligent machines" (p. 2). Scherer (2015) builds on this term by referring to machines that are capable of performing tasks that if performed by a human would be said to require intelligence. The High-Level Expert Group on AI provides, as previously mentioned, a more detailed task description that distinguishes between interpretation, reasoning, and decision making.

Any organization that would like to minimize risks involved in the deployment of AI technologies must define exactly what the organization regulates; in other words, it must find a comprehensive, enterprise-wide definition of the distinctions between AI and human intelligence (Scherer, 2015; Gurkaynak et al., 2016). As consensus remains weak regarding the distinction between AI and human intelligence, it is important to revisit these definitions and ensure that the meanings of *AI* and *intelligence* reflect current changes and challenges. Thus, the aforementioned working definitions of the High-Level Expert Group on AI will be used to avoid ambiguity.

2.3 Ethical Guidelines and Organizational Culture

As the range of ethical values applying to AI constantly shifts, there are situations in which there is a debate regarding the substantive benefits of an AI development to individuals or society. For example, if a client asks to improve their facial recognition software for lethal autonomous weapons systems, then a company could argue for this project by seeing its involvement as contributing to the invention of a murder weapon, but such an improvement could also be interpreted as a way of saving innocent lives. Google's TensorFlow AI systems, for instance, caused controversy among their employees by providing computer vision algorithms to the US military:

> Around twelve Google employees are believed to have left their jobs because of the company's decision to provide [AI] to the Pentagon as part of the US military's Project Maven [. . .] It is expected to develop [AI] capable of sifting through vast quantities of aerial imagery and recognizing objects of interest. This, many Google staff fear, puts the project on a slippery slope towards the weaponization of AI, as the technology could easily be applied to improve the efficacy of drone strikes, for example (Murison, 2018).

As views in ethical debates could differ, it is not possible to objectively define subjective terms such as *ethically justified* or *desired*. This means that companies cannot simply convey ethics values that correspond with each employee's moral compass; transferring an ethical framework in such a way that employees are motivated to follow its guidelines would be difficult, if not impossible.

To ensure that employees are motivated to follow guidelines maintained by employers, it is therefore important to identify primary factors that can motivate employees to act according to an organization's standards, values, and beliefs. The findings of many researchers, including Meglino and Ravlin (1998), Branson (2008), and Sullivan, Sullivan, and Buffton (2001), indicate that the culture of an organization could avoid any possible ambiguity regarding subjective terms by explicitly transferring their standards and values with which employees can personally identify. In this way, values can serve as tools to live by their rules: "values directly affect behavior in that they encourage individuals to act in accordance with the [company's values]" (p. 378). Sullivan et al. (2001) complement this: "It is important for individuals to become aware of their own values as well as the organization's values. The organization's values will signal its direction and the individuals' values will provide their motivation and increase their focus and contribution" (p. 250). Thus, culture, when defined as a mixture of values, beliefs and explanations of behavior could strongly influence an employee's performance (Awadh & Saad, 2013; Hoogervorst, van der Flier & Koopman, 2004). In this way, it could act as behavioral guidance.

If an organization desires to encourage its employees to act in accordance with the company's framework, then the company's culture should be conveyed clearly. According to Branson (2008) and Sullivan et al. (2001), an organization's culture provides a reference point in each individual's decision making to decide whether actions are appropriate and relevant. Sullivan et al. explain that there is less to guide actions in these values-led organizations: "There is less need for detailed procedures [. . .] individuals are [empowered] to make decisions within the framework provided by the organization's values" (p. 249).

As mutual debates regarding ethical decisions could occur, it is thus important to provide developers with tools that direct them to recognize risks and opportunities throughout the entire project. In order to avoid ambiguity as much as possible, this study focuses on a specific company's culture and its standards, values, and beliefs. The European Commission's guidelines act as a filter for the key values of this company so that employees have a clear reference point in each decision when developing, deploying, or utilizing AI.

2.4 Key Findings

F2.1[1] Any organization that would like to minimize risks involved in the deployment of AI technologies must define exactly what the organization regulates; it must find a comprehensive definition for both AI and human intelligence.

F2.2 Owing to a continuous shift in what is considered to be intelligence, which makes it difficult to devise a timeless definition for AI, consensus remains weak on this term (see F2.1).

F2.3 As a result of F2.2, there is a lack of comprehensive technological regulation, as any definition likely would be over-inclusive or under-inclusive.

F2.4 An organizational culture could avoid ambiguity regarding subjective terms by explicitly transferring standards and values with which employees can personally identify.

F2.5 The company's standards and values (see F2.4) provide a reference point in each individual's decision making to decide whether actions are appropriate and relevant.

3 Part I: Sketching the Current Situation

Obstacles that were identified through the literature research show that both the changeability in AI-enabled technologies and differences in personal standards and values make it difficult to compose a comprehensive technological regulation. To determine how these factors play a role in the challenge of applying AI responsibly, an empirical qualitative

[1] This is an abbreviation for "Finding 2.1, Finding 2.2," etc.

research was conducted to map current problems and possible solutions within the AI fields of different IT companies. These interviews occurred before the guidelines were published, allowing the stakeholders to have independent views regarding the introduction of guidelines within their companies. By mapping these problems and solutions, the purpose of the first part of the study was to gain insight into the current ethical codes of different organizations and their regulations toward the development of AI. Additionally, research sought to answer this question: How do developers currently ensure that their AI developments maximize benefits while minimizing risks?

3.1 Method and Procedures

Part I, called "Sketching the Current Situation," contains 16 exploratory interviews to find out how different IT companies apply a system of ethical governance in AI.

Data Gathering Procedure. Semi-structured interviews were utilized to gather qualitative data. These interviews offered a balance between the focus of a structured ethnographic survey and the flexibility of an open-ended interview, because it encouraged interviewees to share their personal perspectives and experiences without the risk of the answers being socially desirable, meaning that interviewees would answer honestly instead of supplying answers that they expected to be desirable to hear. It also allowed the interviewer space to delve more deeply into particular answers to gather more detailed information.

The interviews were held at various Dutch offices and lasted between 45 and 90 minutes. The interviewees were interviewed separately. The interviews were neither taped nor filmed, but the interviewer recorded their thoughts in order to analyze the answers. Validity of the insights was guaranteed by sending the section 3.2 "Outcomes" and the interviewer's written notes to the interviewees afterward, asking if the comments the interviewer made were justified and complete. Interviewees were granted the opportunity to verify a quote within two weeks.

Subjects of the Study. The interviewees for this subproject were recruited in two ways: either through an email that included a short description of the study or through a personal message via LinkedIn. It is important to mention that neither the interviewees or the interviewer were aware of established guidelines from the European Commission because the interviews occurred from September 2018 through the beginning of January 2019. The European Union's guidelines were published on 18 December 2018. Thus, the interviewees' input is based solely on their own insights.

The interviewees were employed by various corporations that utilize AI, such as IT consultants, financial services suppliers, and healthcare providers. They have been deliberately chosen for their in-depth knowledge of AI, which provided relevant insights to conduct further analysis, and have various roles within the companies, such as managers, developers, and lawyers. It is assumed that they have sufficient knowledge of AI because they have all worked on projects that include AI-enabled technologies. The companies are geographically dispersed.

Data Analysis. The individual interview reports are categorized according to the themes that the interviewee presented. If an aforementioned topic has been discussed, then the results are added to previous comment to identify an aggregate variable. Powerful citations and remarkable statements from every interview were marked to garner an idea of what the interviewees considered to be important. Each literal answer was translated to a higher level, that is, a reasoning or motivation behind the answer (e.g., "If I choose not to develop an AI-enabled technology based on ethical considerations, then our competitor will develop it" is "The influence of a technological rat race"). By extracting these motives, they can be quantified statistically, which made it possible to detect similarities between the interviews.

Based on these similarities, underlying results could be distinguished from each other in two schematic analysis. The insights were placed in a coordinate system to maintain a clear overview. In both models, the Y axis juxtaposed technical and business aspects, while the employee level versus company level is displayed on the X axis. The first model framed all identified problems, and the second model framed all suggested solutions for contracting AI projects. These models were utilized to identify which points were identified by a significant number of respondents and to develop conceptual themes.

3.2 Outcomes

The 16 exploratory interviews show how developers currently consider ethical debates have been gained, including identified problems and proposed solutions. This section highlights points that were identified by a significant number of respondents. They have been deliberately highlighted per conceptual theme to prevent that we unintentionally define frameworks through personal biases. The solutions that are presented have been proposed by one or more respondents. See Table 2 for main findings, illustrated in an overview of the identified problems and solutions. A more thorough description of the functions within the interviewees' companies can be found in Appendix I.

Table 2. Identified problems (P) and suggested solutions (S) in the order in which they were discussed in text. These are subdivided in five conceptual themes.

Number	Description
	Providing stability in an ethical framework
P1A	Time-bound concepts, shifting ethical values, and unknown innovations ensure that a set of ethical rules cannot be truly comprehensive.
P1B	A detailed framework must be constantly adapted to changes.
S1	Basic principles could serve as a stable framework to enlarge individual awareness without intricately discussing every aspect.
	Regulating the technological rat race
P2A	Ethical choices to take on assignments are influenced by the desire to utilize new technologies, which can play a key role in competitiveness.
P2B	By taking on assignments, companies partially retain control and the direction of technological innovations.
S2	A sector-wide ethical platform could oblige employees to look objectively at the long-term effects of AI on society by working together competitively.
	Installation of an ethics committee
P3A	No one can give a sophisticated answer to what is ethically responsible as this assessment is partly subjective.
P3B	An ethics committee could have its own interests, which would affect an objective assessment.
P3C	Employees could relinquish their own responsibility to continue thinking about the possible consequences of an AI development.
S3A	An external party that has no economic interests in this regard, and truly does have a say, could assess the need for an AI development.
S3B	An internal ethics committee could play a supporting role in this assessment as long as it is not seen as a gatekeeper. Employees maintain responsibility to think for themselves.
	Rewarding ethical considerations
P4A	Developers are not being judged on making ethical choices as this is not within their scope of functions.
P4B	Employees pass their ethical responsibilities to the client because they assume that the client has already made a moral assessment.
P4C	Ethical objection is perceived by some as omission, so employees do not feel entirely free to take action.
S4A	Enable anonymous questioning of ethical issues with a committee.
S4B	Dynamic contracts could prevent the cancellation of a project from being seen as a default.
S4C	Individual assessment should be entered as part of the code of conduct.
	Reducing conflicts of interest between sales and developers
P5A	Individuals' moral compasses differ due to diverse responsibilities or goals that employees desire to achieve.
P5A	Key Performance Indicators (KPIs) are an incentive to overpromise.
S5	Sales must obtain more knowledge regarding the technical aspects, or developers need permission to voice opinions in undertaking a project.

The Gifts of Imperfection

When asked why composing ethical guidelines is difficult, such drafting committees generally answer that it is complicated to describe accurately which requirements technological developments must meet. The assessment of the ethical aspect of an assignment is more complex than a constant set of extensive rules due to various factors: time-bound definitions, shifting ethical values, and unknown innovations. Interviewee 5 said the following about these shifts: "Developments that aren't considered as ethically justified today, are probably [going to be] justified within five years. How should developers judge developments that are time-bound and focus on the present instead of long-term results?" This view was supplemented by Interviewee 1 and Interviewee 2 who indicated that an unclear vocabulary is the cause of this problem: "What are we talking about when we talk about AI? Without a clear definition, situations cannot be clearly identified, and explicit IT-guidelines cannot be established." Interviewees indicated that this ambiguity can lead developers to be unaware of the content or impact of an assignment, whereby a questionable assignment can be subconsciously accepted.

For this reason, Interviewee 12 provided basic principles that serve as a stable framework to test whether an AI development is desired: "The legal framework tries to create stability, which means that it slowly responds to technological developments. Companies try to incorporate this when they write internal regulations. I think that this unknown direction cannot be sufficiently predicted, and it is smarter to open an ethical debate by excluding a number of points within a framework." This working method was encouraged by two other interviewees who indicated that frameworks can serve as ethical platforms. According to Interviewee 1, a company must set a minimum level of safety for execution of the assignment: "By offering a questionnaire, it becomes clear which assignments are approved by the company's values." Interviewee 13 added: "Yet if an employee monitors all possible consequences and has drawn up a plan for the undesirable effects, they should be given freedom and confidence by their employer to develop everything that is allowed according to law and their own moral standards." Interviewee 3 indicated, however, that a company must be careful that a checklist is not decisive: "A questionnaire is too limited. The culture must be sufficient for the employee to know which assignments are approved by the company."

It can be concluded that ethical frameworks and principles do not exclude unwanted developments. The time-bound concepts, shifting ethical values, and unknown innovations ensure that a set of ethical rules cannot be truly comprehensive. A stable framework can, however, enlarge individual awareness without intricately discussing every aspect; the framework could include questions such as the following: How do you apply this specific AI development? To what extent could developers utilize and further develop AI technologies that may include ethical objections? These concrete tools facilitate a collective discussion.

The Danger of an Ethical Review Committee

A variety of answers were submitted to the question of whether an ethical review committee can support developers in assessing advancements. Interviewees 2 and 14 indicated that there is a need for a sector-wide ethical platform. According to Interviewee 14, a technological rat race currently ensures that companies observe their competitors' behavior to determine if an AI development is being implemented: "It is difficult to determine how to deal with the latest developments. If we do not seize opportunities that another company will seize, we are lagging behind." Interviewee 2 defended this behavior: "By taking on such assignments, you partly retain control and thereby direct the direction of technological innovation. Otherwise, there is a chance that a non-expert will take care of this." A committee with clear ethical standards and values can oblige companies to work together competitively and thereby objectively examine the long-term effects of AI on society.

Yet it appears that not everyone is in favor of an ethics committee. For example, Interviewee 12 indicated that an ethics committee is no more than an instrument: "The danger of a committee is that people let go of their own responsibility to keep thinking about the consequences. They could think that something is ethically justified if the committee approves their request." Interviewee 4 had an equal response: "As a government, but also as a company, you cannot give a sophisticated answer to what is ethically desired. This cannot be tested. An authority could give certain guidelines on how people should act, but they cannot tell what is allowed. Everyone should think about this individually."

In addition to the fear of losing individual responsibility, two respondents (Interviewees 2 and 7) indicated that an internal ethics committee can also have its own interests. Interviewee 7, for example, indicated that companies can encounter pitfalls of risk management: "An assessment can be made based on risk analysis, leaving room of unethical assignments. Because the chance of discovery or occurrence of the unethical behavior of the system is deemed low, the company takes the risk and moves forward." Interviewee 2 agreed and indicated that companies' financial interests can hinder them from making informed choices. The prevention of such entanglements can be handled by an external party that has no economic interests in this regard. A caveat is that these committees have little say, and they are often installed for the sake of appearances.

Although there are both supporters and opponents for the installation of ethics committees, every respondent believed that every person has a moral duty to consider whether a system is ethically and legally acceptable in addition to considering the impact of the development on society. Ethics committees can play a supporting role, but they cannot function as gatekeepers. This arrangement requires everyone to be keen to ponder the ethical side effects for themselves. In addition, ethics committees can only provide objective advice if self-interest is absent.

Ethical Considerations Are Not Rewarded

It appears that it is not simple to refuse an assignment due to ethical considerations. From the responses of several interviewees (5, 6, 7, and 12), this study has revealed that developers do not look beyond the manner in which technical possibilities are optimized, because this is not within the scope of their functions. This does not mean that they proceed without any ethical feeling, but ethical questions (such as, Do I think this development is useful for society?) are not necessarily asked. Interviewee 8 mentioned that: "AI has an enormous impact, because organic growth is lacking in comparison with robotization. Users are part of the change within this process. In some situations, developers take little too little responsibility in their 'big bang effect' on society or within a company." Interviewee 7 offered an explanation for this behavior: "Developers are not being judged on making ethical choices. In addition, there is no support for refusing [an] assignment, because it is not the role of a technician to be a whistleblower."

When two managers from the same company (Interviewees 3 and 5) were asked regarding their experiences with this problem, they indicated that there is no clear protocol for developers' actions. One manager responded, "An employee has the right to say 'No' to an assignment that goes against their personal values. An employee knows what the company stands for, but they can still get into a conflicting situation. It is too easy to say that 'Everything should be fine' within an assignment without looking critically at the project." The second manager said, "If you are an employee, there should be no restrictions on the execution of assignments. If there is something that is in conflict with an employee's personal values, another employee could possibly take over the project." These responses illustrate that an individual has the right to refuse an assignment, but that it is common that colleagues, with their individual principles and values, are able to approve the project. Ethical objections are, therefore, quickly seen as personal objections.

Interviewee 7 indicated that the profit-seeking aspect can be a cause for overlooking ethical considerations: "We often assume that the customer has made a moral assessment before they give us an order. With that, the responsibility no longer seems to lie with us." Nevertheless, a board member of the same company as Interviewee 5 indicated that this is not entirely true: "The motivation of the customer must always be considered to assess whether the project is suitable for implementation."

The responses demonstrate that employees are expected to critically examine assignments. An underlying problem that accompanies such examination is that there is no ethical standard; therefore, every assessment is partially subjective. It is a challenge to weigh the various interests. Also, because colleagues have the opportunity to adopt assignments, ethical objections can be perceived by some as failing, causing employees to feel a hesitation to take action. Subsequently, three potential solutions have been provided for the lack of rewarding ethical considerations. Interviewee 7 indicated that the possibility of raising these issues anonymously with a committee would be helpful. Alternately, dynamic contracts could be devised, whereby the cancellation of a project is not seen as default. Finally, Interviewee 4 stated that this individual assessment should be entered as part of the code of conduct: "Making ethical considerations should not be a reward. It is a basic principle."

Reducing Conflicts of Interest Between Sales and Developers

In product development, there must be individual awareness concerning social consequences as well as technological consequences of a project. Interviewees 2, 3, 5, 6, and 11 indicated that the gap between different departments, therefore, must be narrowed, and their communication must be enlarged to prevent mismatched expectations and empty promises to clients. Interviewee 3 mentioned that he noticed that the need *not to participate* is growing: "From a technical point of view, almost anything is possible, but it is important to look carefully at the drivers for a project." According to Interviewee 6, individuals' moral compasses can differ due to different responsibilities or personal goals: "A sales manager might want to sell a product for the KPIs, while a developer examines whether it is technically a challenge to participate in the same project. The ethical assessment can thus be approached from a different perspective." This was confirmed by Interviewee 2, who indicated that conflicts of interest are a major part of the ethical discussion: "KPIs are an incentive to overpromise. For this reason, our sales team must obtain more knowledge about the technical aspects before they accept an assignment. Another solution is that developers get more say in taking on a project, like Google does."

4 Part II: The Current Method in Light of the European Union Guidelines

Part I of this study shows that there were different factors that could be taken into account while composing ethical guidelines. For example, it was stated that making ethical choices is not within developers' scope of functions and that they pass their ethical responsibilities to the client. By asking developers to utilize the assessment list created by the European Commission, we could gain an understanding of how their current working method corresponds with the key components of this original list. Also, this study explored which areas need attention to ensure that developers are aware of and trained in trustworthy AI. This subproject is called "The Current Method in Light of the EU Guidelines."

4.1 Method and Procedures

To make a proper translation to company-specific guidelines, the research needed to focus on a specific case. The second part of this study, called "The Current Method in Light of the

European Union Guidelines," was conducted to gain insight into the extent to which developers currently consider the European Commission's guidelines.

Data Gathering Procedure. This sub-study consists of a combination of two techniques. First, developers who were involved in this specific case were asked to complete the original version of the European Commission's assessment list. As a result, it was possible to observe similarities and differences between the current working methods of the developers and the requirements that the European Commission wishes to consider. If a question was not clear or not applicable, the participant could answer "Unclear" or "Not applicable."

Subsequently, semi-structured interviews were utilized to obtain structured answers to key components. This also allowed the interviewer space to delve more deeply into particular answers to gather more precise information. The interviews took an average of 28 minutes to complete and were held at the developers' offices or via Skype. The developers were interviewed individually. The interviews were taped in order to later analyze the answers. Validity of the insights was guaranteed by sending the section 4.2 "Outcomes" and the transcript to interviewees afterward, asking if the comments the interviewer made were justified and complete. Interviewees were granted the opportunity to delete, extract, or rectify any comment within two weeks.

Subjects of the Study. To avoid any form of ambiguity with the first study, the interviewees from the second study are referred to as *participants*. The participants are 3 out of 10 project members from the particular use case (see Figure 1). The participants have different functions (i.e., software developer and AI developer), but all are employed by Capgemini. Capgemini is a multinational professional services and business consulting corporation that provides IT services to its clients and has over 200,000 employees worldwide in over 40 countries. Besides these 3 participants, the scrum master of the Applied Innovation Exchange of Capgemini is interviewed as Participant 4. All participants were contacted via personal messages on WhatsApp or via email.

The European Commission's assessment list. The questionnaire consists of 65 individual questions that are subdivided into ten topics. The distribution of the questions is as follows: `ACC` (7), `DG` (5), `DFA` (5), `GAA` (6), `ND` (5), `RFP` (4), `RFHA` (4), `R` (16), `S` (6), and `T` (7). See Table 1 on page 6 for the fully written-out requirements. Participants were free to respond in the way they considered necessary. If they were unable to provide a specific answer, they were asked to reply with "Not applicable" or "I don't understand the question, because [specific reason]." It took, on average, 65 minutes to complete the assessment list. The participants utilized an average of 685 words to answer. Participant 1 utilized 925 words (14.23 words per question), Participant 2 utilized 231 words (3.55 words per question), and Participant 3 utilized 899 words (13.83 words per question). Participant 4 was not asked to complete the assessment list as the scrum master was asked to determine whether the assessment could be applied within Capgemini, based on these results.

In this assessment list, 36 questions were closed-ended, and 29 questions were open-ended. A question was considered closed-ended if the participant could reply with nothing more than "yes" or "no," "true" or "false," or a name. This subdivision is furnished in Appendix IV.

Data Analysis. The answers to the European Commission's assessment list were placed side by side to observe differences and similarities. The questions that none of the three developers (i.e., Participants 1, 2, and 3) could answer were highlighted. These participants were asked, by means of personal messages via email, what they found unclear about those particular questions. The questions from the initial list that the participants answered unanimously were subsequently taken out. The participants were asked what they found clear about those questions. Then the differences in answers were highlighted in the other questions. A number of notable answers were presented to ask the participants for clarity.

After the interviews, each interview was fully transcribed. Powerful citations and typical comments were marked to gather an idea of what the interviewees found important. First, all problems were highlighted in yellow, and all suggested solutions that the developers could implement on their own were highlighted in orange. Thereafter, all other remarkable comments were highlighted in blue. Second, each literal answer was translated to a reasoning or motivation behind the answer by writing insights in the margins of the reports (e.g., "I mainly stick to my own standards and values" is coded as being self-evident). Various key findings have been unearthed from these insights.

4.2 Use Case: Project the AI Experience

The Ministry of Justice and Security contracted Capgemini to build an interactive AI demonstration that showcases how an AI system trains and learns and then demonstrates what it has learned. Their goal was to make AI understandable for the average visitor during the Innovation Congress "What's Next?" on 20 November 2018.

Motivation to utilize this case. There are several reasons why this project was chosen as the use case. First, mutual debates regarding ethical decisions could occur concerning the ultimate goal of this project. Does an AI development that recognizes deviant behavior contribute to improving society? How could developers, for example, decide fairly which behaviors should be implemented as deviant or wrong? Second, Capgemini has a strong corporate culture and core values that are transferred through employee onboarding programs and e-learning courses. As an ethical company, Capgemini could have a starting point for their ethical framework and the emphasis could be translating the European ethics guidelines to the developers. Third, the use case was finished before the guidelines were published; therefore, the developers could only ensure that the risks were minimized according to their own standards, which could coincidentally overlap with the ethics guidelines of the European Commission.

Development of the AI demonstration. By explaining the workings of AI in their basic forms, the developers thought this technology would be more accessible to many people. Therefore, they decided to program their AI demonstration to recognize six different poses: waving, hitting someone, having a cup of coffee, calling, giving the "thumbs up" sign, and pointing.

To do this successfully, the demonstration was divided in two parts: AI training and showcasing the AI's accuracy. The first part of the demonstration functioned purely as the gathering of training data, and the second part was utilized for testing and exhibiting the accuracy of the model. Participants in the AI demonstration could test its accuracy by posing in one of the predetermined ways and seeing if this was correctly interpreted by the AI (see Figure 2). Since the AI model trained itself during set intervals of 15 minutes, its accuracy would increase during the congress.

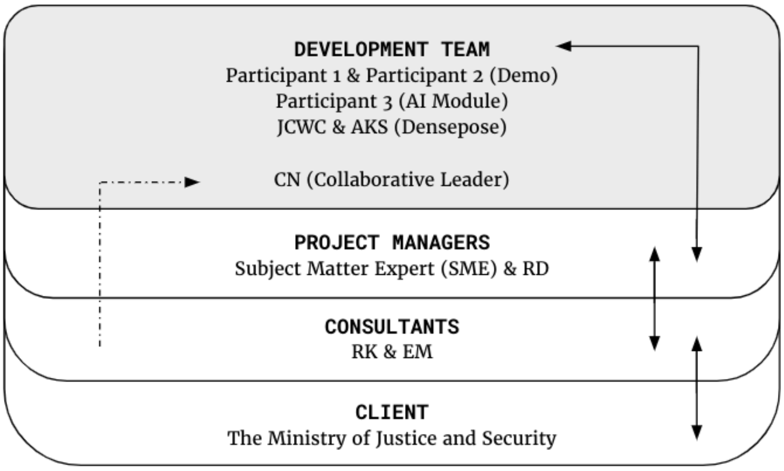
DEVELOPMENT TEAM
Participant 1 & Participant 2 (Demo)
Participant 3 (AI Module)
JCWC & AKS (Densepose)
CN (Collaborative Leader)
PROJECT MANAGERS
Subject Matter Expert (SME) & RD
CONSULTANTS
RK & EM
CLIENT
The Ministry of Justice and Security

Figure 1. Overview of the project members' roles in Project the AI Experience as explained by Participant 4. As scrum master of Capgemini's Applied Innovation Exchange, participant 4 has an overarching role in all its projects.

Technical deficits within the module. It appeared that the AI demonstration was not successful due to a lack of training data among other reasons. First, there were insufficient images to train the model. The developers implemented as many high-resolution images as were available (approximately 50 images per pose). By mirroring and transforming the images as well as creating a personal database of images by taking different photos of employees, the database was expanded to 300 images per pose. It appeared that this amount was insufficient to train a module of this scale. Also, the AI demonstration was not able to discriminate between similar poses if the picture resolution was inadequate. The waving, thumbs up, and calling poses were more or less the same, which led to an equal prediction for these stances. All poses that the AI had to recognize featured the hand raised near the head. Finally, a bias led to discrimination as female attendees were outnumbered by male attendees who were wearing a suit. Because of a small data set, the AI module recognized incorrect patterns. The AI associated women with 2 poses, while it recognized more difference between the poses of men.

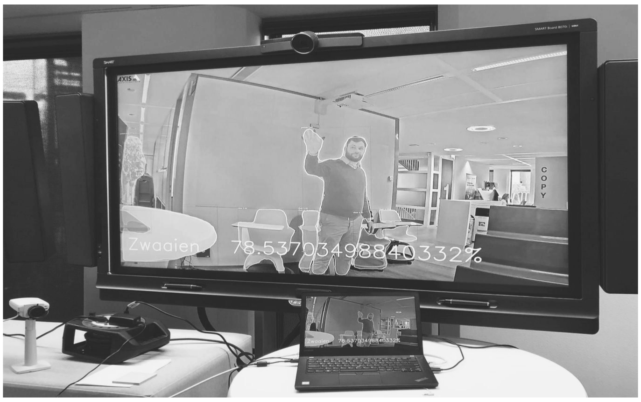

Figure 2. The AI module detects that a person is waving with an accuracy of 78.54%.

Future opportunities for Project the AI Experience. The ultimate goal of the developers was to use AI to recognize poses in public spaces, which can be expanded to identify other expressions, such as aggressive behavior. This technology could be implemented as a security measure for crimes that are fairly predictable and manageable in certain environments.

For example, ProRail has considered behavioral recognition to be used to prevent person-to-train collisions and to reduce the amount of suicides. According to data from ProRail, the company responsible for the Dutch railway network infrastructure, 215 out of 248 suicides attempted by stepping in front of a train in 2017 were fatal (Hermanides, 2018). Hermanides (2018) indicates that the behavior of potential suicidal people can be recognized by their doubtful behavior and by exploring the environment. In that situation, they lean forward, walk back and forth or smoke. An AI system could support or supplement camera images, surveillance employees, or predictive profilers in the future by foretelling or recognizing undesirable behavior, thus leading to early intervention in public spaces. Such a system involves identifying certain behaviors that, if deviant and recognizably wrong, can easily be predicted and controlled.

4.3 Outcomes

Part II of this study provided insight regarding the extent to which developers' current working methods correspond with the key components of the European guidelines and explore which areas need additional attention. This section is subdivided into three different portions. First, a statistical background concerning results is provided. Second, the general design of the questionnaire is mentioned, whereby the clarity of questions, deficits, and suggestions are discussed. The shortcomings and suggestions that are submitted have been proposed by one or more participants. Subsequently, specific attention was paid to the answers furnished by the participants in the assessment list: Which aspects did they consider during the implementation of the project? Why have some aspects not been put in context? Moreover, similarities and differences in their responses are discussed with Participant 4, who is the scrum master of this division. Finally, key findings are presented. See appendices for the participants' answers (II), a transcript of interviews (III), and the distribution of the developers' answers (IV).

Statistical View on Outcomes

Within the development team, one or more participants were unable to provide a complete answer to 17 of the 65 questions. The developers furnished equivalent answers to 23 questions, which means that the answers were not contradictory. Different answers, meaning answers in which the participants contradicted each other, were supplied for 25 questions. For example, answers such as "I am not responsible" and "The project manager is responsible" were seen as contradictory. An overview of these results is illustrated in Table 3. If the answers were incomplete, then the developers either did not respond at all, or they indicated that the meaning of the question was unclear. These questions can be found in Table 5.

The results demonstrate that equal answers were provided for 17.24% of the open-ended questions compared to 50.00% of the closed-ended questions (See Table 4). Conflicting answers were provided for 62.06% of the open-ended questions compared to 19.44% of the closed-ended questions. The developers were unable to provide complete answers on 20.68% of the open-ended questions compared to 30.55% of the closed-ended questions.

Table 3. The distribution of developers' answers in percentages (Participants 1, 2, and 3). See Appendix IV for a more thorough overview.

Requirement	Equivalent answer	Different answer	Incomplete answer
ACC	28.57	42.86	28.57
DG	20.00	60.00	20.00
DFA	20.00	20.00	60.00
GAA	83.33	16.67	-
ND	40.00	20.00	40.00
RFP	25.00	75.00	-
RFHA	75.00	-	25.00
R	25.00	56.25	18.75
S	16.67	33.33	50.00
T	42.86	28.57	28.57

Table 4. The results of the open-ended questions are displayed on the left side. On the right side, the results on the closed-ended questions are displayed. The results are divided per requirement.

Req.	Equivalent	Different	Incomplete	Req.	Equivalent	Different	Incomplete
ACC		2/2		ACC	2/5	1/5	2/5
DG		3/3		DG	1/2		1/2
DFA		1/3	2/3	DFA	1/2		1/2
GAA	3/4	1/4		GAA	2/2		
ND			1/1	ND	2/4	1/4	1/4
RFP		2/2		RFP	1/2	1/2	
RFHA				RFHA	3/4		1/4
R	2/11	7/11	2/11	R	2/5	2/5	1/5
S		1/2	1/2	S	1/4	1/4	2/4
T		1/1		T	3/6	1/6	2/6
SUM	5 (17.24%)	18 (62.06%)	6 (20.70%)	SUM	18 (50.00%)	7 (19.44%)	11 (30.56%)

This Is Not Our Responsibility

There is a preference for handling the guidelines prior to beginning a project so that it becomes clear which values the company desires to meet. Participant 3 indicated that this allows developers to understand in advance the ways in which values can be implemented according to the company's standard. This was also discernable in a response from Participant 1: "In that way, project members cannot realize that something was not ethically responsible after the developed product has been used." Participant 2 deviated slightly from the other two answers by indicating that he considered the guidelines only necessary if the output of a project is utilized at a social level. In that case, "it needs to be clear what a developer could, should and is expected to pay attention to."

It appears that some developers do not desire to take responsibility for a large part of the issues covered in the questionnaire. Participant 2 indicated that project members have different tasks and responsibilities, and therefore, some questions do not concern all the developers: "As you can see, the last 2 or 3 questions in the DG concern an oversight mechanism and legislation that don't have anything to do with the code. The project leader is the person responsible for ensuring that guidelines are pursued by everyone [. . .] I want to know which aspects I have to take into account and where I should focus on. That should be purely about the code and algorithms." Participant 3 also mentioned that he did not want to take ultimate responsibility for the project: "Developers are expected to think along with the design of a system [. . .] They can partly contribute to [the design of a project]. [. . .] It is the product owner's responsibility to ensure that the system complies with the ethical guidelines [. . .] not the developers. Frankly, I do not want to take this responsibility." Participant 1 deviated from this attitude toward work. His answers appeared to indicate a need to answer ethical questions in groups. He also indicated that it is important that every project member must complete the assessment list at least once before insignificant factors are removed in later iterations.

Based on the answers to the questionnaire, this study confirms that the developers do not see the bigger picture and are uninformed regarding organizational structure. For example, it is unclear who serves as primary contact and who is responsible if problems arise. It appears to be unclear who is ultimately responsible for DG with the developers, the Subject Matter Expert (SME) and Capgemini as an entire organization seen as responsible entities. None of the three developers could identify who is responsible to ensure that AI systems are properly governed (GAA.23); therefore, Participant 3 referred to the SME. As a final example, three different answers were provided for question ACC.3. Participant 1 indicated that reports can be made to developers, while Participants 2 and 3 indicated that there is no formal process for handling reports.

Table 5. The questions that have not been fully answered. Gray indicates that the question's meaning was unclear; black indicates that the participant did not respond or that it was not possible to explain what was unclear.

Requirement	Question	P1	P2	P3
ACC.2	Are the skills and knowledge present in order to assume the responsibility?	▒		
ACC.5	Is an (external) auditing of the AI system foreseen?	▒		
DG.8	Is proper governance of data and process ensured?			▒
DFA.13	Is the system equitable in use?			▒
DFA.16	What definition(s) of fairness is (are) applicable in the context of the system being developed and/or deployed?			▒
DFA.17	For each measure of fairness applicable, how is it measured and assured?			▒
ND.24	What are the sources of decision variability that occur in the same execution conditions? Does such variability affect fundamental rights or ethical principles? How is it measured?	█		█
ND.25	Is it clear, and is it clearly communicated, to whom or to what group issues related to discrimination can be raised, especially when these are raised by users of, or others affected by, the AI system?	▒		
RFHA.36	Do users have the facility to interrogate algorithmic decisions in order to fully understand their purpose, provenance, the data relied on, etc.?			▒
R.37	What are the forms of attack to which the AI system is vulnerable? Which of these forms of attack can be mitigated?			▒
R.40	Are the algorithms utilized tested with regard to their reproducibility? Are reproducibility conditions under control? In which specific and sensitive contexts is it necessary to utilize a different approach?	█		
R.51	In case of unacceptable impact, have thresholds and governance for the above scenarios been defined to trigger alternative or fall-back plans?			█
S.54	For each form of safety to be considered, how is it measured and assured?			█
S.55	Have the potential safety risks of (other) foreseeable uses of technology, including accidental or malicious misuse thereof, been identified?		▒	
S.56	Is information provided in case of a risk for human physical integrity?	█	█	█
T.64	Is the nature of the product or technology and the potential risks or perceived risks (e.g., around biases) thereof communicated in a way that intended users, third parties, and the public can access and understand?		█	
T.65	Is a traceability mechanism in place to make my AI system auditable, particularly in critical situations? This entails documentation of: [. . .].		█	

Make Meaningless Words Concrete

There are several reasons developers could not fully respond to all questions, including problems with question clarity and overly complicated answers. Participant 1, for example, explained that it is difficult to answer whether the skills and knowledge are present in order to assume the responsibility because it is not clear which skills and knowledge a developer needs according to Capgemini. Participant 3 experienced the same problem with the word *etc.* in question `RFHA.36`: "I have difficulty understanding the question because the word *etc.* is not meaningful. What else should I take into account to give users the facility to interrogate algorithmic decisions?" Another problem encountered in the questionnaire was the subjectivity and ambiguity of specific words. Terms such as *proper, fairness, accuracy, safety* and *sources of decision variability* can be interpreted in various ways, making it difficult to provide a confident answer to `DG.8`, `DFA.16`, `ND.24`, `R.44`, and `S.53`. Participant 3 referenced the question "Is proper governance of data and process ensured?" (`DG.8`) as an example: "I do not understand the term 'governance of process' and the word 'proper' is extremely subjective." If this explanation is compared with his answer to the question, then it is striking that he attempted to guess what was meant by the question: "I am not certain if I understand this question correctly. If this has to do with the GDPR laws, we have taken some measures." Participant 2 left all answers blank, choosing not to explain any issues with the questions. However, he stated in the interview that he did not understand the purpose of some questions or the relevance of considering a specific aspect: "It has to be clear why something is needed." This was also mentioned by Participant 3 in response to question `R.37`: "I am not aware of all the different ways the AI system could be vulnerable. I would say one of the risks is [. . .]."

A number of solutions are provided to resolve these vague questions. All developers, for example, indicated that they prefer to utilize a checklist, which would make general questions more interpretable because it would be simpler to observe the extent to which a system is compliant with ethical guidelines. "Developers can answer with 'Yes' or 'No' and draw a percentage from the amount of answers that have been given. I think that would be nice, because [we] only have to run a list and check if it is applicable," said Participant 3. He supplemented this by saying that such a statement should be on a scale from 1 to 5 so that developers can indicate that something is partially completed. For example, Participant 1 indicated that there could be a checklist for question `ACC.2` with skills and knowledge needed. Participant 3 indicated the same for question `ACC.1`: "This could be simplified by putting names under each requirement and an explanation that the project member (or entire team) should think about." Second, a further clarification of the purpose of the question should be implemented. Participant 2 explained that it would be nice to have the possibility to click on an example of a case. This could enlarge awareness of the potential impact on society: "What are the consequences of a development? I believe that developers currently think that this is not a big deal and that it is [another's] role. I think it is crucial that they know something about it as well." It could, therefore, reduce the aforementioned uncertainty of a question. Participant 3 said: "I think that it would help to have an example of a case, so it is clear what is exactly meant by each question. The usage of words is complicated and a bit tricky, so I am not always sure what is really expected from me." It is, however, important to keep the questions as generic as possible because detailed guidelines constantly need to be adjusted.

Key Findings

F4.1 Developers prefer closed-ended questions as subjective and ambiguous words can be interpreted in various ways, making it difficult to supply a confident answer.

F4.2 Developers prefer to handle the guidelines prior to a project so that the items to which they could, should, and are expected to pay attention become clear.

F4.3 Developers currently do not see the bigger picture and are uninformed regarding the organizational structure, such as who serves as primary contact or who is responsible if problems arise.

F4.4 Developers need to understand the purpose of a question or its relevance to consider a specific aspect.

A Scrum Master's Opinion on the Project Members' Roles

To determine whether the key findings were project-specific or could apply more widely within Capgemini, the scrum master of the Applied Innovation Exchange was interviewed. The scrum master indicated that the projects are innovative and dynamic, making it difficult to discuss all possible problems beforehand: "There are some factors that you have to adjust in the middle of development, because you cannot [. . .] exclude everything in advance. That is why you will also have to pay attention during the process."

Yet the scrum master also prefers to handle the guidelines prior to a project (see F4.3). "I prefer to determine [factors such as intended use] in advance, so that the project members can clearly see whether [their project] is successful or not [. . .] you try to take this into account," she said. She emphasized that every project member is responsible for the process, so they are expected to think along with the design of a system. She explained that a lack of understanding the purpose of a question or its relevance (F4.4) can arise due to a lack of communication between the project manager and the development team. For example, a project manager can choose to involve developers in their communication with the client or not to pass on details.

A closed-ended checklist provides more clarity so that project members can indicate whether they have what is needed for a project (see F4.1). A disadvantage is that a generic checklist is not comprehensive for a specific project. A proposed solution is to supply a number of focus points that project members must consider. Developers can draw up their own checklists, based on these points, which they can discuss in a group context. The problem with F4.1, however, remains in every checklist: "Questions must be clearly defined because developers can be addressed otherwise that an AI-development is incomplete when the [subjective, ambiguous or incomplete words] get meaning."

For Capgemini the checklist system would work if the relevance of such a questionnaire was clear: "If there is no need for this [. . .] they will not do anything with it." By completing the list prior to beginning a project, members can make their assessments more specific and determine their own needs in the list (F4.6). This generic list can act as a guide, but the fact remains that not everything can be thought out in advance with dynamic projects: "I would make sure that [developers] are not deposit[ed] with too much documentation. Get started and [they] will gradually see if something is wrong."

5 Recommendations Regarding Company-Specific Guidelines for Capgemini

To determine how these guidelines can be utilized effectively, it is equally important to determine potential difficulties in the assessment list. The results reveal that there are various hindrances within the AI field. First of all, the subjectivity in defining ethical concepts and the unique view that everyone possesses on an ethical dilemma make it difficult to devise a predetermined set of rules to make informed choices that adhere to certain principles. Also, the fuzzy front-end of innovation processes[2] as well as the changeability within AI-enabled technologies ensure that an assessment cannot be truly comprehensive. This means that regulations must be constantly adapted to changes. Additionally, composing guidelines while the direction is yet unknown ensures that there can be no stability in such a technological regulation. Moreover, there appears to be tension between gaining freedom and taking responsibility because developers usually do not look beyond the ways in which technical possibilities should be optimized.

These impediments in the questions can be resolved at both the organizational and content levels. In this section, recommendations are presented based on the overarching problems that occurred in both studies.

[2] The early innovation phases "that come before the formal and well structured [product and process development] . . . the activities are often chaotic, unpredictable and unstructured" (Koen et al., 2001, p. 49). So, the unknown unknowns on the innovation process cannot be identified prior to the project.

Example 1. Questions in which meaning is supplied to subjective words to avoid ambiguity. Definitions for *safety* and *fairness* have been obtained from the European Commission's document (Smuha, 2018, p.15-18).

`DFA.16` "What definition(s) of fairness is (are) applicable in the context of the system being developed and/or deployed?"

Further explanation for `DFA.16`:
An AI application is fair if it is user-centric and considers the whole range of human abilities, skills, and requirements to utilize the products or services, regardless of their individual characteristics (such as age, gender, social status, or disability).

`S.53` "What definition(s) of safety is (are) applicable in the context of the system being developed and/or deployed?"

Further explanation for `S.53`:
A system is safe if it minimizes unintended consequences and errors in the operation of the system, such as harming users or the environment. Additionally, unintended consequences should be clarified and assessed, which could then be adapted.

Give meaning to subjective terms to avoid ambiguity. The literature demonstrates that it is difficult to provide a timeless definition for AI because terms of AI are often tied to the ability to perform specific intellectual tasks. This makes it difficult to supply any legal definition for the purposes of regulation. The results of the first part of the study, current problems in the IT field, have also highlighted this problem in two ways. References are made to both shifts in ethical assessments and the lack of a clear vocabulary. The ambiguity in regulation ensures that developers are unaware of the meaning of an assignment.

The European Commission's assessment list contains words that have value judgments which are not precisely defined. For example, the Commission speaks of proper governance of data and process, and definitions of *fairness*, *accuracy* and *safety* are applicable in the context of the system being developed or deployed. The developers indicated that they find it difficult to provide a concrete answer without knowing exactly what is meant by a question. The scrum master indicated that developers find this difficult because people can refer to these indefinite words once they have acquired a meaning. As a result, they prefer to keep questions with ambiguous words unanswered.

To prevent such answers in the future, meaning should be supplied to these subjective words. The literature recommends that the organization's culture is conveyed clearly so that it provides a reference point for each decision, thus ensuring that actions are appropriate. The interviews display, however, that Capgemini's organizational culture is not conveyed clearly and that developers often utilize their own sense of appropriateness to guide their behavior. Developers prefer an enlightening definition of those terms and the option to view an example situation to consider decisions in context. Although both options may be appropriate, an understanding of company culture provides more freedom of assessment by

Example 2. A checklist based on the current method utilized at the Applied Innovation Exchange, applied to a question from the list.

D.8 “Is proper governance of data and process ensured?”		
	YES	NO
If personal or confidential information is utilized, security aspects are taken into account.		
Management procedures have been determined, such as user access, network or system access, and application access.		
The development is able to capture consent for processing through a sufficiently clear privacy notice.		
The privacy notice is written in a child-friendly manner in case minors are utilizing the development.		
All processed data is kept in a Record of Processing, which includes an updating process.		
Any personal data which is held longer than the legally defined retention period is deleted, pseudonymized, or anonymized		
The development contains a functional requirement to give effect to data subject right requests, such as the possibility to delete, extract, or rectify data.		

combining personal standards and values with the company’s principles. A precise definition could, however, provide more clarity as it is more concrete (as illustrated in Example 1).

Add a checklist to questions so it is clear when a developer meets the requirement. By keeping questions as broad as possible, these questions remain less sensitive to time-bound definitions, shifting ethical values, and unknown innovations. Developers find it, however, difficult to answer questions that do not exactly define what should be included. As shown in Example 2, sub-questions can be answered to give a broader perspective to the developers such as ‘Is there a possibility to delete, pseudonymize or anonymize personal data which is held longer than the legally defined retention period?’ and ‘Is the privacy notice written in a child-friendly manner in case minors are utilizing the development?’. By providing a description that gives more context to a question, developers can find more structure in their assessment processes to consider the different requirements, as an interviewee mentioned in the first study. It does not, however, ensure that developers can see the extent to which a system is compliant with ethical guidelines.

The European Commission's assessment list contains broad questions that are not precisely defined. For example, one question asks whether the developers have the skills and knowledge necessary to take on responsibility; another question asks whether the system is GDPR compliant. Participants provide different answers to the questions, namely "Yes," "No," or "As far as I know." By offering a checklist whereby the developers themselves can indicate whether they meet each requirement, it is possible to minimize the number of different answers given by different developers within the same company. The scrum master mentioned, however, that project members must be aware of all project requirements.

Capgemini currently utilizes a questionnaire to determine whether a project is feasible to implement. This includes the topic of data privacy. In Example 2, Capgemini's privacy list has been converted into a checklist that illustrates the way in which a complicated question could be supported. A focal point is that only the primary factors are included to prevent developers from being provided extraneous documentation. A checklist, therefore, is often not comprehensive.

Provide a scale from 1 to 5 with a specific action per answer in closed-ended questions to gain additional insight. Although a checklist provides more clarification in an abstract question, a closed-ended question can also limit the ways in which developers would like to answer. For example, Participant 3 indicated that answers such as "Yes" or "No" are in some cases not exhaustive, and a checklist could be restrictive.

As indicated in the previous recommendation, it can be unclear what the answer "Yes" covers. A scale from 1 to 5 can allow a developer to indicate that something has been partially achieved. The numbers 1 to 5 are not meaningful without adding values. It is therefore recommended to supply substance to this scale. By introducing different degrees of actions, a developer can indicate that an item is slightly finished as well as discover additional items that could further optimize a development. An illustration is provided in Example 3.

Provide an example situation with each question to illustrate its relevance to consider a specific aspect. The results demonstrate that shifting responsibility is an overarching problem. Although the developers indicated that it is important to consider the possible impact of an AI development on society, they also indicated that it is the product owner's responsibility to ensure that the system complies with ethical guidelines. This belief was also reflected in the first interviews, in which interviewees disclosed that employees pass their ethical responsibilities to the client because they assume that the client has already made a moral assessment. With this in mind, it appears that only a portion of developers feel responsible for the product development as well as its consequences.

Although participants shirk responsibility, the responses also reveal ignorance in some questions. Participant 2, for example, provided a conflicting answer when he said that there is a need for the ability to click on an example of a case that could explain the consequences of a development. This more comprehensive understanding could enlarge developers' awareness of the possible impact on society. The scrum master explained that developers will not do anything with a question if there is no need to answer it. By making their assessment more specific, they can determine their own needs in the list.

The participants indicated that they, therefore, need an example situation that illustrates the possible impact of an AI development. This could increase the relevance of considering a specific aspect. This can be found in Example 4.

Example 3. A scale from 1 to 5 with a specific action per answer in a closed-ended question.

`DFA.14` "Does the system accommodate a wide range of individual preferences and abilities?"

Individual characteristics are not taken into account.	It does not have a one-size-fits-all approach, as it can be utilized regardless of age, gender, or social status.	It also considers accessibility and usage for people with disabilities.	It provides choice in methods of usage and provides adaptability to the user's pace.	All aspects are included and users have the option to propose changes in the system.

Example 4. Utilizing an example to illustrate the relevance of a specific aspect to consider.

`ND.26` "Is a strategy in place to avoid creating or reinforcing bias in data and in algorithms?"
Relevance of `ND.26`: "Discrimination in an AI context can occur unintentionally due to [. . .] problems with data such as bias, [or] incompleteness [. . .] Machine learning algorithms identify patterns or regularities in data and will therefore also follow the patterns resulting from biased and/or incomplete data sets. An incomplete data set may not reflect the target group it is intended to represent" (Smuha, 2018, p. 16).
Example for `ND.26`: PredPol, an American software company, predicted where crime would occur based on previous reports. The algorithm chose how police officers were distributed between different locations. If an arrest took place, then other officers were sent to the same location. This reinforced the chance that more crime would be discovered there: "That means the software ends up overestimating the crime rate in one neighborhood, without taking into account the possibility that more crime is observed there simply because more officers have been sent there—like a computerized version of confirmation bias" (Reynolds, 2017). Due to an incorrect feedback loop and a data set based on reports that already contained a bias, police behavior in regard to racial prejudice was increasingly reinforced by a one-time iteration.

6 Discussion

In this section, both fundamental limitations and resource-driven limitations of the study are considered. Subsequently, recommendations for further research are made.

6.1 Limitations of the Study

There are some fundamental problems within this research, making it difficult to devise an objective assessment list for AI-enabled technologies. These are described in the first subsection of the limitations of this study. Also, difficulties that arose during the study are described in the second subsection.

6.1.1 Fundamental Limitations

Ethics is relative to the person, environment, and situation. This not only means that every developer has a unique view on an ethical dilemma or a specific choice in the development process, but also that a person's ethical vision can change per project. This changeability of the subjective assessment makes it difficult to compose guidelines: How could developers decide fairly which AI developments are justified and desired?

It is not possible to objectively define subjective or ethical concepts, so companies must convey their ethical values in such a way that developers are motivated to follow any technological regulation. Particular guidelines are a means to indicate which actions must be taken to ensure that a development is built in the ideal fashion. Because ethical dilemmas require choosing between alternatives that must be evaluated as either right or wrong, however, there is no possibility to devise guidelines that contain the complete truth. Subjectivity appears, therefore, to be one of the fundamental difficulties in any technological regulation.

Despite this subjectivity, a predetermined set of rules must be composed to make informed choices that adhere to certain principles with the aim of developing trustworthy AI. Regardless of the quality of this research, a pre-established list cannot be truly comprehensive, even if there is consensus on the design of the guidelines and which criteria are applied. This is due to the fuzzy front-end of the innovation process as well as the dynamics within these technologies which make it impossible to determine in advance what will or can happen in the future. This ensures that regulations must be constantly adapted to changes. Also, composing guidelines while the direction is yet unknown ensures that there can be no stability in such a technological regulation.

These two problems require conflicting solutions. In the case of ambiguous regulations and vocabulary, concrete guidelines can ensure that subjectivity is less important because objective measures provide greater clarity. The results demonstrate that this is desired by some developers: subjective and ambiguous words can be interpreted in various ways, making it difficult to provide a confident answer. However, a dynamic project requires abstract guidelines that can be implemented in various ways. As Participant 4 mentioned: "There are some factors that you have to adjust in the middle of development, because you cannot [. . .] exclude everything in advance. That is why you will also have to pay attention during the process." Through abstract guidelines, an assessment list can be supplemented if a previously unknown change takes place.

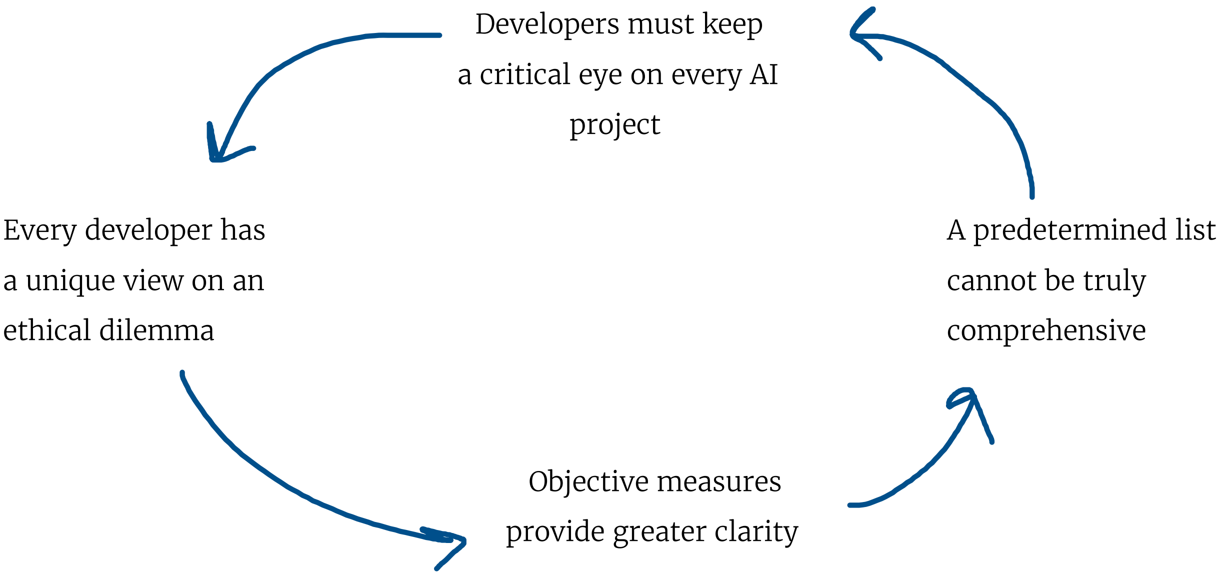
Developers must keep a critical eye on every AI project
A predetermined list cannot be truly comprehensive
Objective measures provide greater clarity
Every developer has a unique view on an ethical dilemma

Figure 3. Subjectivity cannot be excluded in an AI assessment list.

The drafting of a specific set of ethical rules that are accepted by all developers remains, therefore, a contentious issue complicated by fundamental limitations. Although an assessment list can be given to developers to steer them in a certain direction, it does not necessarily exclude unethical developments. This means that developers must continually keep a critical eye on every AI project. Looking beyond the guidelines, however, leads to more subjective assessments that were initially corrected with the drafting of a technological regulation (shown in Figure 3).

6.1.2 Resource-Driven Limitations

This study had a number of resource-driven limitations that must be cited. As little research had been conducted into devising ethical guidelines regarding AI for companies, a start had to be made from a certain point. This means that this exploratory study lacks replication.

Part I: Sketching the Current Situation. The first part of the study occurred between September 2018 and January 2019. Initially, the aim was to establish ethics guidelines for Capgemini whereby their developers learned to utilize this technology in a responsible manner. To be able to impart full advice, we felt it relevant to acquire knowledge concerning both AI and Capgemini's organizational culture. To know the organization inside and out, all research occurred at the Utrecht office of Capgemini.

Through unstructured interviews, an attempt was made to identify the problems that employees encounter to determine where solutions could be offered. With ethical problems and a diversity in individual standards and values, the priority was that the interviewer minimized the chance of bias. The interviewees therefore were allowed freedom to determine issues themselves, whereby the interview could be partially directed by asking preset questions, such as "Do you believe you need ethical guidance on the implementation and operationalization of AI?". It is important to note that influencing by the interviewer lurks in qualitative research (e.g., inadvertently directing the conversation through personal interest in specific issues or responding positively to a particular answer by nodding in satisfaction). Additionally, it is possible that the interviewees did not feel sufficiently free to raise all problems they experience. The lack of anonymity present in this type of research can lead to socially desirable answers being provided. Moreover, an interviewee can only be challenged by the interviewer to consider new perspectives because the conversations occurred on an individual basis.

To avoid a tendency to offer socially desirable answers, these conversations were not recorded; only notes were written with the conversations. Although the validity of the insights is guaranteed by allowing interviewees the opportunity to review their comments, a recording could have prevented an unfair distribution in the results.

As a final point, it was important that there was a diverse group of interviewees to be able to map problems in all organizational layers. This research began with five employees who hold managerial positions within the IT industry (Interviewees 2, 3, 4, 5, and 6). Following these initial discussions, other interviewees were approached within equal or different fields, including education, finance, insurance, and health care. The key findings can therefore apply to problems occurring throughout the entire Dutch IT industry. However, this study worked with a small group of participants, making it possible that not all problems have been identified or that a number of problems have not been sufficiently identified.

Part II: The Current Method in Light of the European Union Guidelines. During the completion of these exploratory interviews, the European Commission published its first version of the guidelines for AI. Since this research is based on the vision of 54 stakeholders, their advice is the basis for devising company-specific guidelines. This eliminated a number of ethical bottlenecks that emerged in the discussions, such as determining which requirements an AI must meet to be ethically responsible or taking a step backward when it is ethically justified. Additionally, it allowed more stability and objectivity to the size of this research.

To gain an understanding of the ways in which the developers' current working methods corresponded with the key components of the original list, the participants from the use case were asked to assess the list only after their project was finished. The advantage of this strategy is that the developers acted solely on their own principles and the protocol of Capgemini, so a clear separation could be made. The disadvantage, however, is that the project members saw less applicable items in the usage of this assessment list as their project was already complete. For this reason, the other project members within Project the AI Experience were unable to complete the list because they had other priorities. Consequently, the research contains the perspective of the development team only, while many answers refer to the Subject Matter Expert. This one-sided vision is a limitation in this study. Future research should include both multiple visions and another case study that utilizes the assessment list in all steps.

Given these points, if this study is repeated, then the results will not necessarily be the same, because every project is unique. The reader should remember that this study is based on a number of opinions that ultimately make no truth. This study, however, focuses on the underlying problems in the European Commission's assessment list and the ways in which this transition can be made for Capgemini specifically. It is therefore important that additional research is conducted because this study is merely the beginning of research into composing ethical guidelines regarding AI for IT companies.

6.2 Recommendations for Further Research

Because changes in ethics and AI occur every week, this study focuses solely on the guidelines of the European Commission and Capgemini's aim to provide developers with tools. This particular assessment list is nevertheless only a selection of the various forms of ethical testing related to AI. For example, Google has appointed an ethics committee of eight members who will meet four times in 2019 to "consider some of Google's most complex challenges that arise under [their] AI principles, like facial recognition and fairness in machine learning" (Walker, 2019). This committee is an extension of the concrete standards that Google had previously established, such as being socially beneficial, to ensure that the development of AI would be built and tested for safety as well as being accountable to people (Pichai, 2018). IBM has also introduced new trust and transparency capabilities for AI on the IBM cloud to:

> Provide a level of transparency, auditability and explainability by logging every individual transaction throughout a model's operational life [. . .] and bridge the gap between data scientists, developers and business users within an organization providing them visibility into what's happening in their AI systems (Puri, 2018).

Although this study demonstrates that there is a strong need for specific guidelines and that the responsibility to continue thinking ethically is shifted if someone else is liable, it cannot be claimed that abstract guidelines or the installation of an ethics committee are not at least as effective as composing ethical guidelines. It is therefore important that more research is conducted regarding different forms of ethical steering.

Another subject that remains to be explored is the effectiveness of rewarding developers for their ethical considerations. Some interviewees suggested that an ethical objection could be perceived by some as default and that employees pass their responsibilities to the client or colleagues (P4B and F4.4). An individual assessment as part of the code of conduct is suggested as a solution (S4C). As this study is based on the assumption that ethical guidelines are preferred, the developers within Capgemini mentioned their preferences with regard to design and organization. It has, however, not yet been investigated how this introduction of guidelines will work in practice. Adaptations motivated by this new type of accountability could thus be questioned. Different governance regimes, such as hierarchical, motivational, and political mechanisms could, therefore, be approached to explore which design could be pursued in a way that is sensitive to social and emerging ethical concerns. Hartswood, Grimpe, Jirotka and Andersom (2014), for example, mention in their paper *Ethical Governance of Smart Society* that "it is a very common experience that people are motivated to adjust their practices if they feel that they are being observed or assessed" (p. 25). Guihot et al. (2017) mention, however, that "self-regulation [. . .] works best where there is some imminent threat of state-based penalty for noncompliance" (p. 50). Further research could therefore explore the effectiveness of a hierarchical governance regime, including codes of conduct, monitoring, and penalties, compared to a motivational governance regime with the aim of applying a system of ethical governance in AI and robotics.

A final relatively narrow but important question that was identified after the data collection is this: what impact does the phrasing of a question have on its interpretation? The outcomes reveal conflicting results. Although developers were able to answer more open-ended questions than closed-ended questions, more matching answers were provided for closed-ended questions than for open-ended questions.

7 Conclusion

This study sought to answer the following research question: How could developers include the European Commission's ethical guidelines to add to the striving toward trustworthy AI? To this end, a qualitative study was conducted to map current problems and possible solutions within the AI fields of different companies and to gain an understanding of the ways in which developers' current working methods correspond with the key components of the European assessment list.

To determine how these guidelines can be applied effectively, we believed that it was equally important to define potential difficulties in this assessment list. The results demonstrate that there were various difficulties within the AI field. First of all, the subjectivity in defining ethical concepts and the unique view that each person possesses on ethical dilemmas make it difficult to devise a predetermined set of rules to make informed choices that adhere to certain principles. Also, the fuzzy front-end of innovation processes as well as the changeability within AI-enabled technologies ensure that an assessment cannot be truly comprehensive. This means that regulations must be constantly adapted to account for technology changes. Additionally, composing directives when the direction of a project is yet unknown ensures that there can be no stability in such a technological regulation. Finally, there appeared to be tension between gaining freedom and taking responsibility because developers usually do not look beyond how technical possibilities should be optimized.

Based on these overarching difficulties, recommendations have been provided to improve the assessment list to ensure that Capgemini's developers are aware of and trained in trustworthy AI. For a start, ambiguity must be avoided by supplying meaning to subjective terms. This ensures that developers truly understand the purpose of a particular question. Next, an example situation belonging to each question could illustrate the relevance of developers considering a specific aspect. Besides the usage of an example, a checklist could be added, because developers find it difficult to answer questions that do not exactly define what should be included. Although developers could provide additional insight into the extent to which a system is compliant with ethics guidelines, answers such as "Yes" or "No" may not be exhaustive. A weighted scale with a specific description of an action per answer could, therefore, provide additional insight into these questions. Finally, there must be an option in the fuzzy front-end innovation to add unknown aspects later in the development process.

At the organizational level, it is recommended that all points be discussed in a group context, so that all employees think along with the design of a system. In this way, ethical governance becomes part of the organizational culture. This could solve the lack of a reward policy in which employees pass responsibilities to their colleagues or clients when they believe it is not within their scope of duty. Also, an interdisciplinary discussion could lead to different insights that may otherwise be overlooked in an individual assignment. Furthermore, it is important to limit the amount of questions to the essentials so developers are not overloaded with documentation.

This qualitative study has demonstrated that subjectivity, time-bound concepts, and unknown innovations ensure that an AI assessment list cannot be comprehensive as it does not exclude unethical developments. This means that developers must maintain a critical eye on every project. If the European Commission's ethical guidelines are, however, formulated in a clearer way in response to the aforementioned recommendations, then this list could be a start to increase both the sense of responsibility and the ethical awareness of developers.

8 References

Awadh, A. M., & Saad, A. M. (2013). Impact of organizational culture on employee performance. *International Review of Management and Business Research*, *2*(1), 168-175.

Boddington, P. (2017). *Towards a code of ethics for artificial intelligence*. Cham, Switzerland: Springer.

Branson, C. M. (2008). Achieving organisational change through values alignment. *Journal of Educational Administration*, *46*(3), 376-395.

Guihot, M., Matthew, A. F., & Suzor, N. P. (2017). Nudging robots: Innovative solutions to regulate artificial intelligence. *Vand. J. Ent. & Tech. L.*, *20*, 385.

Gurkaynak, G., Yilmaz, I., & Haksever, G. (2016). Stifling artificial intelligence: Human perils. *Computer Law & Security Review*, *32*(5), 749-758.

Harbers, M. (2018). Verstand erbij: Verantwoord ontwerp van toepassingen met kunstmatige intelligentie. Rotterdam, Netherlands: Hogeschool Rotterdam Uitgeverij.

Hartswood, M., Grimpe, B., Jirotka, M., & Anderson, S. (2014). Towards the ethical governance of smart society. In *Social Collective Intelligence* (pp. 3-30). Springer, Cham.

Hoogervorst, J., van der Flier, H., & Koopman, P. (2004). Implicit communication in organisations: The impact of culture, structure and management practices on employee behaviour. *Journal of managerial psychology*, *19*(3), 288-311.

Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*, *51*(8), 56-59.

Koen, P., Ajamian, G., Burkart, R., Clamen, A., Davidson, J., D'Amore, R., ... & Karol, R. (2001). Providing clarity and a common language to the "fuzzy front end". *Research-Technology Management*, *44*(2), 46-55.

LaChat, M. R. (1986). Artificial intelligence and ethics: an exercise in the moral imagination. *AI Magazine*, *7*(2), 70.

Luxton, D. D. (2014). Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial intelligence in medicine*, *62*(1), 1-10.

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, *90*, 46-60.

McCarthy, J. (1998). What is artificial intelligence?

Meglino, B. M., & Ravlin, E. C. (1998). Individual values in organizations: Concepts, controversies, and research. *Journal of management*, *24*(3), 351-389.

Scherer, M. U. (2015). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harv. JL & Tech.*, *29*, 353.

Sullivan, W., Sullivan, R., & Buffton, B. (2001). Aligning individual and organisational values to support change. *Journal of Change Management*, *2*(3), 247-254.

Torresen, J. (2018). A review of future and ethical perspectives of robotics and AI. *Frontiers in Robotics and AI*, *4*, 75.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433-460.

Winfield, A. F., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180085.

Non-Academic References

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Davenport, T. & Katyal, V. (2018, December 6). Every Leader's Guide to the Ethics of AI. Retrieved from https://sloanreview.mit.edu/article/every-leaders-guide-to-the-ethics-of-ai/

Hermanides, E. (2018, April 13). Suïcide op het spoor: het went nooit. Retrieved from https://www.trouw.nl/samenleving/suicide-op-het-spoor-het-went-nooit~a8d9e10c/

KLM. (2017, Dec 19). KLM zet volgende stap in gebruik kunstmatige intelligentie op sociale media. Retrieved from https://nieuws.klm.com/klm-zet-volgende-stap-in-gebruik-kunstmatige-intelligentie-op-sociale-media/.

Murison, M. (2018, May 15). Google faces rebellion over military AI projects. Retrieved from https://internetofbusiness.com/google-pushback-military-ai-projects/

Pichai, S. (2018, Jun 7). AI at Google: our principles. Retrieved from https://www.blog.google/technology/ai/ai-principles/

Puri, R. (2018, Sep 19). It's time to start breaking open the black box of AI. Retrieved from https://www.ibm.com/blogs/watson/2018/09/trust-transparency-ai/

Reynolds, M. (2017, Oct 4). Biased policing is made worse by errors in pre-crime algorithms. Retrieved from https://www.newscientist.com/article/mg23631464-300-biased-policing-is-made-worse-by-errors-in-pre-crime-algorithms/

Smuha (2018, December 18). *Ethics Guidelines for Trustworthy AI: working document for stakeholders' consultation.* Retrieved from https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_draft_ethics_guidelines_18_december.pdf

Verhagen, L. (2018, Aug 31). *Hoe de politie oude zaken nieuw leven inblaast met slimme computers.* Retrieved from https://www.volkskrant.nl/wetenschap/hoe-de-politie-oude-zaken-nieuw-leven-inblaast-met-slimme-computers~bd5aeaec/.

Walker, K. (2019, Mar 26). *An external advisory council to help advance the responsible development of AI.* Retrieved from https://www.blog.google/technology/ai/external-advisory-council-help-advance-responsible-development-ai/