How do software companies deal with artificial intelligence ethics? A gap analysis

Vakkuri, Ville

ACM
2022-06

Vakkuri , V , Kemell , K-K , Tolvanen , J , Jantunen , M , Halme , E & Abrahamson , P 2022 , How do software companies deal with artificial intelligence ethics? A gap analysis . in M Staron , C Berger , J Simmonds & R Prikladnicki (eds) , Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022 . ACM , New York þÿ , pp. 100 109 , International Conference on Evaluation and Assessmen Engineering , Gothenburg , Sweden , 13/06/2022 . https://doi.org/10.1145/3530019.3530030

http://hdl.handle.net/10138/353975
https://doi.org/10.1145/3530019.3530030

unspecified
publishedVersion

How Do Software Companies Deal with Artificial Intelligence Ethics? A Gap Analysis

Ville Vakkuri
University of Jyväskylä
Jyväskylä, Finland
ville.vakkuri@jyu.fi

Kai-Kristian Kemell
University of Helsinki
Helsinki, Finland
kai-kristian.kemell@helsinki.fi

Joel Tolvanen
University of Jyväskylä
Jyväskylä, Finland
joel@shareway.fi

Marianna Jantunen
University of Jyväskylä
Jyväskylä, Finland
marianna.s.p.jantunen@jyu.fi

Erika Halme
University of Jyväskylä
Jyväskylä, Finland
erika.a.halme@jyu.fi

Pekka Abrahamsson
University of Jyväskylä
Jyväskylä, Finland
pekka.abrahamsson@jyu.fi

ABSTRACT

The public and academic discussion on Artificial Intelligence (AI) ethics is accelerating and the general public is becoming more aware AI ethics issues such as data privacy in these systems. To guide ethical development of AI systems, governmental and institutional actors, as well as companies, have drafted various guidelines for ethical AI. Though these guidelines are becoming increasingly common, they have been criticized for a lack of impact on industrial practice. There seems to be a gap between research and practice in the area, though its exact nature remains unknown. In this paper, we present a gap analysis of the current state of the art by comparing practices of 39 companies that work with AI systems to the seven key requirements for trustworthy AI presented in the "The Ethics Guidelines for Trustworthy Artificial Intelligence". The key finding of this paper is that there is indeed notable gap between AI ethics guidelines and practice. Especially practices considering the novel requirements for software development, requirements of societal and environmental well-being and diversity, nondiscrimination and fairness were not tackled by companies.

CCS CONCEPTS

• **Software and its engineering** → **Software development process management**; • **Computing methodologies** → **Artificial intelligence**; • **Social and professional topics** → *Codes of ethics.*

KEYWORDS

Artificial Intelligence Ethics, Responsible AI, Gap Analysis

ACM Reference Format:
Ville Vakkuri, Kai-Kristian Kemell, Joel Tolvanen, Marianna Jantunen, Erika Halme, and Pekka Abrahamsson. 2022. How Do Software Companies Deal with Artificial Intelligence Ethics? A Gap Analysis. In *The International Conference on Evaluation and Assessment in Software Engineering 2022 (EASE 2022), June 13–15, 2022, Gothenburg, Sweden.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3530019.3530030

EASE 2022, June 13–15, 2022, Gothenburg, Sweden

ACM ISBN 978-1-4503-9613-4/22/06…$15.00
https://doi.org/10.1145/3530019.3530030

1 INTRODUCTION

The discussion on Artificial Intelligence (AI) ethics has been accelerating [9]. What was once largely an academic discussion focused on hypothetical future scenarios has gained traction globally as said scenarios are becoming reality following advances in AI. With ethical issues in these systems, e.g., various data handling issues, making the global headlines, companies have become aware of the importance of ethical consideration during development as well.

As a result, AI ethics discussion has recently begun to also focus on what to do to tackle various AI ethics issues, aside from defining what are key AI ethical principles (i.e. which issues to tackle). Various different types of organizations including governments and companies have begun to devise AI ethics guidelines to help tackle these issues [9]. However, guidelines alone are likely not sufficient in developing ethical systems, as discussed by Mittelstadt [13] in relation to the ACM ethical guidelines in IT in general. In an existing paper, we also argue that this is the case in the area of AI ethics as well, based on industrial survey data [19]. Indeed, implementing abstract principles highlighted by such guidelines can be difficult for companies if there is no actionable advice for doing so.

With the discussion on AI ethics accelerating, the current state of the industry remains a relevant question. With all these guidelines, and other resources such as tools [14], now at their disposal, what are companies doing to implement AI ethics? We have provided a quantitative overview of the current state of industry in another paper [19], highlighting a gap in the area. However, given the quantitative approach, the exact nature of the gap remains unclear. In this paper, we go look into this gap in more detail using a qualitative research approach.

Better understanding this gap in this area is important. By understanding what issues companies currently face in implementing AI ethics, we can better develop solutions and take action to help them tackle these issues. We are currently aware that a gap exists, but the exact pain points remain largely unknown.

With data from 39 companies, we seek to understand what the gap between research and practice in the area is in AI ethics. Using the EU Ethics Guidelines for Trustworthy AI as a ethical framework [6], we study what practices are used to implement which AI ethics

principles. Specifically, we tackle the following research question: how do companies currently deal with the most common AI ethics principles when developing Artificial Intelligence?

Rest of the paper is structured as follows: section 2 describes the background on AI guidelines and methods for implementing ethics in AI, section 3 explains the used research method and framework, section 4 focuses on the findings whereas section 5 discusses them in further detail, and section 6 concludes the paper by summarizing the findings.

2 BACKGROUND

This section is split into two. In the first subsection, we go over the various AI guidelines devised to help organizations develop ethical AI. In the second one, we discuss methods for developing ethical AI.

2.1 AI Guidelines

As touched upon in the introduction, guidelines have been one of the more prominent tools for implementing AI ethics thus far. Governments, standardization organizations, and companies alike have devised guidelines intended to help developers implement AI ethics. [9] While the actionability of these guidelines has been criticized [13], they have been widely discussed and a large number of them exists by now.

These AI ethics guidelines have distilled long-standing academic discussion into principles intended to make them more tangible. In a large-scale review of 84 such documents, Jobin et al. [9] found the following principles to be the most commonly discussed one, from most common to least common: (1) Transparency, (2) Justice, fairness and equity, (3) Non-maleficence (i.e. safety, security etc.), (4) Responsibility and accountability, (5) Privacy, (6) Beneficence, (7) Freedom and autonomy, (8) Trust, (9) Sustainability, (9) Dignity, and (10) Solidarity. The first 5, up until privacy, were found in half of the documents they reviewed. Thus, some consensus in these documents exists, even if the discussed principles were rather varied.

Out of these guidelines, perhaps the most high-profile ones are the IEEE Ethically Aligned Design guidelines (EAD) [5] and the EU Ethics Guidelines for Trustworthy AI [6]. EAD, written by scientists, presents a more academic research focused set of guidelines with 250+ pages of content. The EAD guidelines extensively discuss the research behind the featured principles in detail, focusing on presenting the background behind these principles. EAD also discusses policy and law regarding the context. The principles discussed in EAD are human rights, well-being, data agency, effectiveness, transparency, accountability, awareness of misuse, and competence [5].

The EU guidelines [6], similarly start by defining what is trustworthy AI and what principles produce trustworthiness in AI. These guidelines discuss the following principles: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination, and fairness, (6) environmental and social well-being, and (7) accountability. Perhaps the most important feature of these guidelines, however, is the section dedicated for *assessing* whether a system is trustworthy. The assessment list provides a sheet with questions used to assess whether each principle, e.g. transparency, is tackled and to what extent in the system. This increases the actionability of the guidelines, as these questions can conversely be potentially used to guide development.

Nonetheless, the principle-focused guidelines have been criticized for lacking in actionability [13]. While they are high abstraction level tools that can help companies develop ethical AI systems, they seldom touch upon development in practice. This is where methods step in – or should step in, as few currently exist for AI ethics as we discuss next.

2.2 AI Ethics Methods

Methods in software development are guidelines that describe how work should be carried out. Methods consist of a collection of techniques, which in turn describe how certain, more atomic, work tasks should be carried out [17]. What the Information Systems discipline refers to as techniques are in modern SE literature generally referred to as (development) practices.

Morley et al. [14] conducted an extensive review of methods, tools, and research that could be useful in translating AI ethics principles into practice. Based on the review, most methods in the area seem to be currently focused on smaller subsets of the development process, such as how to manage machine learning. Though useful in their specific contexts, they do not consider the AI development process as a whole.

In addition to those covered by Morley et al. [14], more recent existing research proposes some relevant artefacts. For example, we have proposed the ECCOLA method for implementing AI ethics in a recent paper [22], which builds on the ethical principles seen in the various AI ethics guidelines. In a similar fashion, Canca [2] proposes a tool they refer to as *The Box*, intended to help organizations "determine the relevant ethical concerns and weigh applicable instrumental principles to determine how to best satisfy core principles by substituting or supporting one instrumental principle with another." The Box can, in this fashion, help organizations implement these principles in practice.

With companies largely left to rely on theory-focused guidelines, the current state of practice becomes an interesting question. We have already argued in past studies that companies do not seem to have any formal ways of tackling AI ethics issues and hardly discuss using AI ethics guidelines to do so either (e.g., [19] [21]). However, while these studies identify a gap, our understanding of the exact nature of this gap is still lacking. What exactly is the gap between research and practice in the area?

3 RESEARCH METHODOLOGY & STUDY DESIGN

In this section, we discuss how the study was carried out in three subsections. First, we present the research framework that was used to guide data collection. Then, we discuss the data collection in detail. Finally, we discuss our data analysis method.

3.1 Research Framework

As the research framework for this study, we utilized a framework described in an existing paper here [20]. Based on the academic AI ethics discussion, we chose some of the most prominent principles

to devise a research framework for the survey (which is discussed in the next section) that was used to collect data for this study.

The principles chosen for the framework were Accountability [4][5], Responsibility [4], Transparency [4][5][6][18], and Predictability. These have all been widely discussed principles present in various guidelines, including the EU guidelines used to analyze the data in this study (as we discuss in detail later in this section). Jobin et al. [9] also point out the prevalence of these principles in their extensive review of AI ethics guidelines published after this framework was devised. At the time, these principles were selected based on their prominence in academic discussion, as well as the emphasis placed on them in some of the more high-profile AI ethics guidelines such as those of IEEE [5].

Transparency and predictability are two somewhat connected principles. Ultimately, both are focused on ensuring that we understand how AI systems work. We should e.g., be able to know how the data is collected and handled within the system, and how the system itself makes decisions [6]. This also includes the development process itself, where decisions and actions should be tracked and traceable. Predictability, on the other hand, is about making sure the system acts in a predictable manner in any given situation - which in turn is easier when transparency is in play.

Accountability, as the name implies, refers to accountability or liability issues in terms of the actions of the AI. E.g., who is liable if or when the system causes issues for someone. Dignum 2017 [4] defines accountability to be the explanation and justification of one's decisions and one's actions to the relevant stakeholders. In this framework, we also consider how accountability issues were handled during development. Transparency is closely linked with accountability, as transparency makes it possible to properly establish accountability based on e.g., decisions made during development.

Finally, for the definition of Responsibility, we refer to the IEEE EAD guidelines [5] and consider responsibility as an attitude or moral obligation to act ethically. I.e. where accountability is externally motivated (e.g., legal responsibility), responsibility is internally motivated (e.g., desire to not cause anyone harm).

In devising the survey, we focused on principles and practical issues related to these principles, so as to avoid asking questions directly about (AI) ethics. Ethics is perceived differently by different respondents and there is hardly any academic (much less industry) consensus on what exactly is AI ethics, and thus discussing it directly with the respondents was something we wished to avoid.

3.2 Data Collection

The objective of this study was to understand how AI ethics principles are implemented in practice, and to what extent. In past studies (e.g., [21]), we have approached similar research topics through qualitative case studies. Here, we do so by means of a survey.

This study is based on survey data on AI ethics issues that received responses from 249 companies. We have published a paper discussing the state of the industry based on this data ("The Current State of Industrial Practice in AI Ethics" [19]). The survey process is also discussed in more detail in that paper. The paper utilized the Likert Scale questions of the survey, which comprised the bulk of the survey. In this paper, on the other hand, we look at the *open-ended* questions of this survey.

The survey was conducted as a structured interview where possible, F2F, or online. When this was not possible, responses were simply collected through the survey as an online survey. Companies included in the survey were software companies in general and not AI companies specifically. 106 of the 249 companies were involved in AI development or had developed AI systems. The respondents who took the survey were working predominantly in either Finnish or US-based companies. More detailed demographic information, in the interest of space in this paper, can be found in the existing paper ([19]).

In this paper, while we look at the open-ended questions of the survey, we also only look at the responses from companies that are (or were) involved in AI development. Given that these questions were optional, not all respondents opted to answer them either. Some of the responses were also lacking in quality (e.g., one-word-responses to all of the questions) and were consequently not included. As a result, for this paper, we ultimately analyzed the responses of 39 companies to the open-ended questions of this survey. These questions are found below in Table 1.

These 39 companies ranged from small (1-9 employees) to large (over 500 employees). The companies included into the study as well as the respective respondents are detailed in Table 2.

In the survey outline, AI ethics was defined via four principles. In alphabetical order, these principles were Accountability, Predictability, Responsibility and Transparency. In the questions, the respondents were asked how their organizational policies and practices take these principles into account when developing AI systems. We discussed this framework in further detail in section "Research framework" above.

3.3 Data Analysis

The data were analyzed through qualitative thematic analysis. We utilized the integrated coding approach described by Cruzes and Dybå [3] to code the data. As a basis for the coding process, we utilized the EU Ethics Guidelines [6] as a framework. However, novel codes were also formed if new, clear concepts that did not fit the framework were identified during the analysis. We anticipated that the framework might not cover all the emerging codes fully and thus the integrated coding approach [3] was selected to support both deductive and inductive coding, keeping the coding process within the context while also supporting the formulation of new codes.

Each coded section of data was assigned an inductive and deductive code. First, each quote was assigned an inductive code as concepts emerged from the data during the analysis. Then, these inductive codes were grouped under deductive codes, which were formulated from the practices described in the EU AI ethics guidelines [6]. Sections of data that did not indicate any practice were assigned the code 'observation'.

In this fashion, the data were coded and arranged into themes to identify central ethical concepts. Based on the identified concepts, we formulated a list of practices. This list of practices, then, was mapped to the key requirements of the EU AI Ethics guidelines [6] to identify which practices contribute to fulfilling each requirement. The goal of the analysis was to understand what practices

Table 1: Open-ended questions

#	Open-ended question
1	What kind of software and for what industry does your organization develop?
2	Is artificial intelligence somehow involved in your software development? If so, how?
3	Who makes the final decisions on software development?
4	If your organization does not utilize a contingency plan for exceptions, how are unexpected situations prepared for?
5	"We have faced issues with our software due to unexpected operation" If you have faced issues with your software, please specify:
6	Do your organization policies consider accountability, responsibility, transparency and/or predictability in your software development? If so, how are they considered?
7	Does the consideration of accountability, responsibility, transparency and/or predictability show in practice in your development processes? If so, how do they show?
8	What do you consider to be the most significant benefits an organization can gain by considering accountability, responsibility, transparency and predictability in their software development?
9	What kind of benefits has your organization gained by considering accountability, responsibility, transparency and predictability in your software development?

Table 2: Respondent company description

Company description	n
Finnish company	23
Finnish public sector	4
Multinational outside EU	3
Other	9

companies use to pursue ethical AI while developing AI, as seen through the lens of the EU AI ethics guidelines.

4 FINDINGS

In this section, we discuss the results of the study. First, in this main section, we provide an overview of the data analysis. Afterwards, each subsection of this section is dedicated to one principle of the EU AI ethics guidelines [6]. In the final subsection we summarize the results.

As we present our results, we summarize our key findings as numbered Primary Empirical Contributions (PECs) for clarity of presentation. We then use these PECs to organize the following discussion section. Moreover, we occasionally utilize direct quotations from the data to liven up the text, but it should be noted that any following conclusions are not based on any single citation alone.

The summary of the results of the coding process can be found in table 5 below. Deductive codes were based on the EU guidelines [6] while inductive ones were formulated based on the data alone.

Thus, based on the thematic analysis, codes of conduct were the most commonly discussed practice. On the other hand, there were also some practices not present in the data at all: ethics and lawfulness by design and diverse and inclusive design teams. To this end, the requirements of societal and environmental well-being, as well as diversity, nondiscrimination, and fairness were not apparent in the data.

PEC1: The requirements of societal and environmental well-being, as well as diversity, nondiscrimination and fairness, were not discussed within the data.

4.1 Human Agency and Oversight

The requirements for human agency and oversight are intended to ensure human autonomy [6]. E.g., AI systems should support human decision-making rather than making decisions for humans. Regulations emerged in the data as one factor supporting the implementation of human agency and oversight. The respondents discussed making sure to adhere to any regulations that might apply to the AI they are developing. Some respondents felt that following regulations was at least what they should do, while some respondents specifically added that following regulations was all they felt they needed to do:

"*We follow regulations, but since our software is not a very risky one, we haven't taken much caution.*"Respondent C18

Aside from regulations, accountability via governance systems is considered to provide an overall oversight of the AI's operation, as well as the company behind the system, and this oversight also covers human oversight. The respondents felt that accountability guidelines contributed towards trustworthiness and ethical AI. To this end, clearly defined responsibilities formed another practice based on the data.

Aside from regulations, explanation methods also contribute towards human agency and oversight. However, only one respondent highlighted the reviewability of their AI models that provided the possibility for human oversight of the system. Thus, these methods did not seem to play a large role in current practice.

PEC2: Companies employ practices that could contribute towards the requirement of Human agency and oversight, but they do not consider the requirements directly, leaving their contribution towards Human Agency and Oversight vague.

Table 3: Companies and respondents for the study

ID	Company size	Role of respondent(s)
C1	1-9	CTO
C2	1-9	Supervisor
C3	1-9	Specialist (creating business development related content/requirements)
C4	1-9	Consultant
C5	1-9	Developer
C6	1-9	Creative director
C7	1-9	Developer, architect
C8	10-49	Sales Director
C9	10-49	Technical service responsible
C10	10-49	Software engineering, product development
C11	10-49	Front end developer
C12	10-49	CEO and product owner/designer
C13	10-49	CEO
C14	10-49	Product owner
C15	50-249	Software designer
C16	50-249	Project manager in delivery projects
C17	50-249	CTO (leading whole RNG incl. SW development)
C18	50-249	AI specialist
C19	50-249	Head of a competence organization
C20	50-249	Software Developer
C21	50-249	Project manager
C22	250-499	Full Stack Developer
C23	250-499	Product manager
C24	250-499	Business line director
C25	500+	N/A
C26	500+	Lead consultant/Architect
C27	500+	Administrator and development specialist
C28	500+	Image preprocessing, training data set validation dataset
C29	500+	Team lead
C30	500+	Integration specialist
C31	500+	Requirements engineer/consultant
C32	500+	Product manager/owner
C33	500+	N/A
C34	500+	Senior testing specialist
C35	500+	Business analyst.
C36	500+	N/A
C37	500+	BI/AI services Area Manager, Project Manager
C38	500+	Project Manager
C39	500+	Senior systems architect

Table 4: EU Guideline Principles and Requirement

Principles	Requirements
(i) Respect for human autonomy	1 Human agency and oversight
(ii) Prevention of harm	2 Technical robustness and safety
(iii) Fairness	3 Privacy and data governance
(iv) Explicability	4 Transparency
	5 Diversity, non-discrimination and fairness
	6 Societal and environmental well-being
	7 Accountability

4.2 Technical Robustness and Safety

Technical robustness and safety refers to minimizing harm that the AI system could cause. To this end, AI systems should be resilient to adversary actions, have procedures to ensure fail-safe operation in unexpected situation, and operate in a predictable and explainable manner [6].

Software architecture plays a key role in ensuring technical robustness. Multiple respondents discussed their role in producing Ethical AI:

"*We focus on finding the right solution, not the easiest*" Respondent C24 "*We always try to select long term architectural solutions*" Respondent C21

Service quality indicators provide different technical metrics such as functionality, performance and reliability, which can be used to assess technical robustness and safety [6]. Three practices explicitly focused on metrics were found in the data. However, these metrics in the data were more concerned with measuring progress in terms of project management, e.g., time spent, rather than assessing the implementation of any ethical principles. I.e., Quality of Service metrics were mostly used internally for project management purposes rather than assessing ethical aspects.

Also relating to software quality, testing and validation practices were discussed in nine cases. Code reviews were the most common practice in this category in the data; having multiple programmers go over the same code was considered beneficial for quality. Moreover, testing overall was considered to be key in producing trustworthiness of the system. Tests were also carried out using real-world scenarios for further validation.

While explanation methods are expected to produce reliability and repeatability, which are requisites for building safe and robust AI systems, these were only discussed by one respondent. This was also the case for audit trails, discussed by the same respondent alone. It would seem that these are not common practices.

Similarly, certification, another way of producing trust to software, was only discussed by one participant who discussed a certified method, SAFe, as a way of maintaining quality. To this end, many of the respondents felt that using agile methods contributes to promoting trust to software in various ways (e.g. transparency to customer, predictability). With few other types of practices discussed in this category, this leads us to suggest the following:

PEC3: Technical robustness and safety are currently realized with existing, common software development methods (e.g. agile) and testing and validation practices (e.g. code reviews).

Table 5: Assigned Codes and Their Occurrences Within the Data

Deductive code	Inductive code	Occurrences
Accountability via governance	Accountability guidelines	2
Accountability via governance	Defined responsibilities	7
Certification	Audits	1
Certification	Certification organizations	1
Codes of conduct	Company policies	4
Codes of conduct	Contract	5
Codes of conduct	Decision-making practices	2
Codes of conduct	Documentation	7
Codes of conduct	Operational guidelines	5
Codes of conduct	Preparation	1
Education and awareness of ethical mindset	Change agency	2
Education and awareness of ethical mindset	Trainings	1
Explanation methods	Reviewable models	1
Explanation methods	Software audit trail	2
Observation		8
Regulation	Following regulations	3
Service quality indicators	Cost tracking	1
Service quality indicators	Time tracking	2
Stakeholder participation	Customer involvement	4
Stakeholder participation	Informed customer	3
Standardization	Agile methods	3
Standardization	Unified processes	6
Testing and validation	Code reviews	3
Testing and validation	Testing	5
Trustworthy architectures	Proper architectural solutions	2

4.3 Privacy and Data Governance

AI systems should ensure privacy, protecting the data in the system from tampering, while also assuring its quality and limiting data access to only those with appropriate reasons to access it [6]. In this category of requirements for ethical AI, regulations such as the General Data Protection Regulation (GDPR) played a notable role. In this regard, too, following regulations was deemed the least companies could do, and in some cases sufficient on its own.

"*All employees are trained to follow organizational policies. Company also follows GDPR strictly.*" Respondent C38

Contracts with customers were discussed in relation to privacy and data handling as well. The companies remarked that data handling related requirements were often defined formally with the customer. However, this would also indicate that data privacy issues are seldom considered with the general public or those whose data is being used in mind. To this end, other types of codes of conduct such as operational guidelines and company policies were also considered by the respondents to contribute towards trustworthy software and thereby to ethical AI.

Some practices such as validation, testing, and architecture could contribute to data governance as well. However, these were not explicitly discussed in relation to it, and thus their role is unlikely to be notable. To this end, we suggest the following: **PEC4**: Privacy and data governance seem to be largely tackled by simply adhering to regulations.

4.4 Transparency

Explanation models are at the core of transparency, and more specifically explainable AI systems. However, only one company, discussed them in their responses in the form of reviewable models and audit trails. While audit trails are existing practices usable in any software engineering project, they were not found in the responses. Audit trail logging could help assess how the system operated in given conditions.

System architecture was more commonly discussed in relation to transparency. Yet, how exactly it was used to support transparency was not detailed.

Transparency is not related to the system alone but the development process as well. In terms of the latter, multiple respondents discussed how their organization kept track of who developed certain parts of the system. While this was done in part simply to keep track of productivity as well, it contributed to transparency. These types of accountability frameworks also serve to establish accountability, as we discuss later in this section.

As briefly touched upon in relation to agile practices earlier, stakeholder participation was also something discussed by multiple respondents in relation to transparency. Stakeholder, in this case, largely referred to just the project customer(s), however, whereas in AI ethics stakeholders are considered more comprehensively. Though the respondents discussed various degrees of customer involvement, customer involvement was often discussed in some way, if only in terms of keeping the customer informed. In cases where stakeholder (customer) communication was mostly one-way, the customers rarely actively contributed to the development process. Agile methods were considered to contribute to transparency in other ways as well:

"*Our aim is to be very transparent for the customer. Our project aim to follow agile methods and as a part of them, the processes like planning sprints and reviewing the developed solutions are very transparent to the customer*" Respondent C26

Also related to development practices, code documentation was an existing common practice that was frequently discussed in relation to transparency. Documentation in terms of project documents related to decision-making were also acknowledged as drivers of transparency by the respondents. Furthermore, the respondents also considered code reviews to produce transparency, even though it was limited to the organization developing the system.

Notably, however, transparency was largely only discussed in relation to customers: **PEC5**: Transparency is primarily perceived as a matter between the software provider and their customer during the development process and it is already realized with various practices, with 10 practices out of 11 having at least a partial contribution to its realization.

The only exception to this observation was transparency mandated by regulations. For example, the GDPR in the context of data handling mandated some transparency outside transparency towards the project customer. Though the respondents that brought it up indicated that their organization followed such regulations, they did not indicate that it would have motivated them to otherwise consider transparency in a wider sense.

4.5 Accountability

Accountability via governance systems is manifested in the industry by defining responsibilities and accountability guidelines. Contracts were typically cited as means of establishing accountability for any system, at the start of a project. One respondent also discussed being responsible for any issue arising during the operational life of the system that could be traced back to the development phase, also covering accountability past the development phase. For the most part, however, accountability past the development phase was left unclear, as the respondents only discussed it in relation to development alone.

The practices that produced transparency were also discussed in relation to accountability by the respondents. For example, code documentation and keeping track of who changes what in the code were also practices that supported accountability. Similarly, documenting decision-making and other project records served to support accountability.

"*These factors have now been added eg to our developmental process, because it has been seen that it is an advantage to increase our customers and our own knowledge about these, especially when facing new challenges with the near future AI technology. It is good to prepare in advance, not after something has happened.*" Respondent C36

As was the case with transparency, accountability, too, was largely considered only in relation to the project customer(s).

PEC6: Accountability is considered to be a matter between the company and its customer(s), and is, in this context, already realized with existing practices such as stakeholder participation and governance systems.

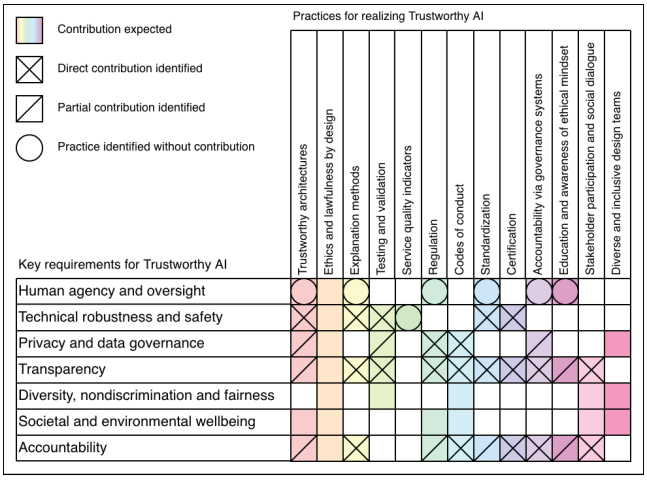

Figure 1: Contributions towards realization of trustworthy AI's requirements

4.6 Summary of Findings

In the Figure 1 , we present our findings in relation to the key requirements for trustworthy AI outlined in the EU AI ethics guidelines [6]. A checked square refers to direct contribution by the practice. A square with one line in it refers to vague or minor contribution. A square with a circle means that a practice was present within the data but no contribution towards realizing that practice was identified. Finally, a coloured square alone means that the practice was expected to contribute towards the requirement, but was not identified in the data. The data in this figure itself presents the seventh and final PEC of this study:

PEC7: Two requirements proposed by the European Commission, (1) ethics and lawfulness by design and (2) diverse and inclusive design teams, were not brought into attention by companies developing AI. Other practices proposed by European Commission are already employed for achieving trustworthiness with contributions of various extent.

5 DISCUSSIONS

To provide a framework for the discussion, we refer to the Primary Empirical Contributions (PECs) highlighted during the analysis. These have been compiled into Table 6 below. We relate each of these findings to existing literature and discuss their implications in this section.

PEC1. The absence of societal and environmental well-being, as well as diversity, nondiscrimination and fairness, can be considered to contradict existing research. Existing research has placed much emphasis on societal and environmental concerns in AI ethics. Furthering human well-being in general is discussed in various AI guidelines [9]. Yet, these types of more general ethical considerations were not present in the data, past some mentions of company goals without concrete practices attached.

In a similar vein, PEC2 highlights another poorly tackled subset of trustworthiness. Human agency and oversight [6] were not directly tackled by any of the involved companies, even though some existing practices could indirectly also address issues related to it to varying extent, leading us to argue that it is largely ignored. Human authority and autonomy over AI systems has historically been an important topic of discussion in AI ethics, including both current systems and still hypothetical future scenarios (e.g., advanced general AI). Delegating tasks that previously required human reasoning to AI systems is one way in which AI systems can reduce human autonomy, and this shift can have unexpected consequences [23]. Education could help mitigate risks in such shifts in authority, and e.g., more widespread programming and software engineering skills could help humans oversee AI systems, as opposed to leaving it to a small group of experts [10].

In terms of technical robustness and safety [6], agile methods, as existing software development approaches, provide means of tackling some of the requirements of trustworthy AI (PEC3). Brock and von Wangenheim [1] also discussed organizational agility and ability to adapt to change as a success factor for AI adoption. Realizing technical robustness as far as reliability and reproducibility go could further be increased by creating AI systems with interpretable and transparent underlying models [16]. The companies in this set of data also selecting the most suitable architecture as opposed to the easiest one, which would indicate that such concerns are tackled at least for software quality reasons.

As progress on AI continues and its impacts grow on a societal level, regulations affecting AI systems are likely to only increase in number. Given how heavily AI systems rely on data, the GDPR has already forced many companies involved with AI to take it into account in developing and operating their systems. Many of the companies in this study as well discussed focusing on adhering to regulations, which they considered sufficient in terms of data-related ethical issues. This situation has been acknowledged by existing studies as well [7].

Together, PECs 5 and 6 form an interesting observation that highlights a clear gap in the area. Whereas AI ethics emphasizes the importance of taking into account the effects of AI on different stakeholders (e.g, in the context of the fairness principle (see e.g. [12] companies seem to still hold a narrow view of stakeholders). In this sense, these observations can be considered to contradict existing studies. For example, the company delivering the AI may only directly interact with the company commissioning that system from them. However, the system, once operational, could then operate using data collected from the general public in various ways, making any subject of this data collection a relevant stakeholder as well. It would seem that accountability and transparency towards such stakeholder groups is left to the customer to tackle – or not tackled at all.

Finally, PEC7 summarizes our results as far as the EU guidelines for trustworthy AI [6] are considered. This conclusion is elaborated on in detail in Figure 1 at the end of the preceding section. Related to PEC1, this PEC highlights areas of AI ethics that seem to not be considered important in the industry currently, from the point of view of the EU guidelines specifically. In the Figure1 , we highlight *which* areas of AI ethics are currently addressed at all and to what extent, based on data from 39 companies.

These findings further our understanding of the current gap between research and practice in the area. AI ethics issues considered important in research are not always considered important

Table 6: Primary empirical contributions formed from the data

PEC	Relation to existing research	Contribution
PEC 1	The requirements of societal and environmental well-being, as well as diversity, nondiscrimination and fairness, were not discussed within the data.	Contradicting, environmental and societal issues were considered as major concerns of AI (e.g. [10]).
PEC 2	Companies employ practices that could contribute towards fulfilling the requirements of Human Agency and Oversight but they do not tackle the requirements directly, leaving the extent to which these requirements are tackled vague.	Contradicting, ensuring human authority was discussed in multiple studies (e.g. [10, 23]).
PEC 3	Technical robustness and safety are currently realized with existing, common software development methods (e.g. agile) and testing and validation practices (e.g. code reviews).	Corresponding with previous research (e.g. [1]).
PEC 4	Privacy and data governance seem to be largely tackled by simply adhering to regulations.	Corresponding with previous studies (e.g. [7]).
PEC 5	Transparency is primarily perceived as a matter between the software provider and their customer during the development process and it is already realized with various practices, with 10 practices out of 11 having at least a partial contribution to its realization.	Contradicting, Transparency should encompass other stakeholders (e.g. people affected by AI supported decisions) than only the paying customer. (e.g. [8, 10, 16].
PEC 6	Accountability is considered to be a matter between the company and its customer(s), and is, in this context, already realized with existing practices such as stakeholder participation and governance systems.	Contradicting, Accountability of AI system's operation should cover other parties in addition to the vendor and customer (e.g. [10, 11, 23]).
PEC 7	Two requirements proposed by the European Commission, (1) ethics and lawfulness by design and (2) diverse and inclusive design teams, were not discussed by companies developing AI. Other practices proposed by European Commission are already employed for achieving trustworthiness with contributions of various extent.	Novel, previous research on how trustworthiness is pursued by companies developing AI systems was not identified.

out on the field. Our findings paint a better picture of *which* issues are currently ignored. Future research can utilize these findings to focus effort towards these less focused on issues (or requirements, in the context of the EU guidelines [6]). We also provide further insights into the extent to which some of the requirements, which are acknowledged, are tackled. The narrow view of the stakeholder discussed in PECs 5 and 6 is one such example.

The primary practical implication of these findings is that companies should use methods or tools for implementing AI ethics. However, methods for this purpose are still scarce. The method we have presented in an existing paper, ECCOLA, is one option in this regard [22].

Regulations and laws set the bare minimum for ethics (PEC4). Though legislation is typically slow to move and lags behind technological progress, increasing numbers of AI-related laws and regulations are being worked on. These include the AI Act of the EU. Tackling AI ethics is becoming increasingly mandatory, and utilizing methods and tools to do so in an organized manner should yield better results and be easier than doing so ad hoc. A method, for example, helps an organization define what AI ethics is, while also providing, ideally, a way to implement it in practice.

5.1 Limitations

The survey used for data collection in this study was devised before the guidelines for trustworthy AI [6] used as a framework in this study were published. As a result, the survey did not directly correspond to all the requirements in the framework. To reduce the effects of this limitation, we utilized the integrated coding approach described by Cruzes and Dybå [3] This also served to address the potential limitation of the data analysis method where the data shapes to reflect these guidelines instead, which can be a pitfall in qualitative research [15].

The data collection method presents another limitation. The structured data collection approach (survey or interview depending on the case) left no possibility for further, elaborating questions. As a result, it was not possible to attain exhaustive descriptions of the practices discussed by the respondents. Moreover, in the cases where data were collected through a survey rather than an interview, the respondents may have been inclined to give less lengthy responses as well to save their own time, further reducing the detail of the responses.

In addition, it should be noted that many of the companies included in the data sample operate in Finland. This may have affected some of the findings. For example, all these companies adhere to the same local and European Union level regulations. Regulations and laws set the bare minimum for ethical consideration. E.g., companies operating in the EU have to adhere to the GDPR. Moreover, with most of the companies being from Finland and the US, and with the data analysis framework being European, this paper overall presents a western point of view on AI ethics.

6 CONCLUSION

In this study, we analyzed survey data from 39 companies developing AI systems or AI-based systems. Through the survey, we sought to understand to what extent they consider AI ethics while working with AI systems. Utilizing the EU guidelines for trustworthy AI

systems [6] we analyzed this data to see what requirements for trustworthy AI are being fulfilled by these companies and to what extent. In doing so, we sought to further our understanding of the gap between research and practice currently existing in the area.

Based on our results, we highlight more specific gaps in the area. These findings, alongside other ones, are elaborated on particularly in Table 6 in the preceding discussion section, as well as Figure 1. We would consider the following two points the key take-aways of this study:

- The industry currently exhibits a narrow view of the stakeholder in the context of AI systems. Stakeholders are considered from the point of view of conventional software engineering projects, and are thus the idea of the stakeholder is largely limited to the customer alone. As AI systems often exert a much larger influence than this that encompasses various stakeholders, this view is narrow in the context of AI. This relates to multiple AI ethics principles such as transparency and accountability.
- When looking at AI ethics through the lens of the EU AI ethics guidelines [6], some categories of requirements for trustworthy AI are currently largely ignored. E.g., companies do not seem to pay much mind to requirements relating to societal and environmental well-being in their software development practices.

REFERENCES

[1] Jürgen Kai-Uwe Brock and Florian Von Wangenheim. 2019. Demystifying AI: What digital transformation leaders can teach you about realistic artificial intelligence. *California Management Review* 61, 4 (2019), 110–134.

[2] Cansu Canca. 2020. Operationalizing AI ethics principles. *Commun. ACM* 63, 12 (2020), 18–21.

[3] D. S. Cruzes and T. Dyba. 2011. Recommended Steps for Thematic Synthesis in Software Engineering. In *2011 International Symposium on Empirical Software Engineering and Measurement*. 275–284.

[4] Virginia Dignum. 2017. Responsible autonomy. *arXiv preprint arXiv:1706.02513* (2017).

[5] Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems 2019. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

[6] Ethics guidelines for trustworthy AI 2019. Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[7] Michael Haenlein and Andreas Kaplan. 2019. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review* 61, 4 (2019), 5–14.

[8] Andreas Holzinger, Peter Kieseberg, Edgar Weippl, and A Min Tjoa. 2018. Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 1–8.

[9] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[10] Andreas Kaplan and Michael Haenlein. 2020. Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons* 63, 1 (2020), 37–50.

[11] Jerry Kaplan. 2016. *Artificial intelligence: What everyone needs to know*. Oxford University Press.

[12] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376445

[13] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* (2019), 1–7.

[14] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2020. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics* 26, 4 (2020), 2141–2168.

[15] Michael D Myers and Michael Newman. 2007. The qualitative interview in IS research: Examining the craft. *Information and organization* 17, 1 (2007), 2–26.

[16] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[17] Juha-Pekka Tolvanen. 1998. *Incremental method engineering with modeling tools : theoretical principles and empirical evidence*. Ph.D. Dissertation. Väitöskirja, Jyväskylä.

[18] Matteo Turilli and Luciano Floridi. 2009. The ethics of information transparency. *Ethics and Information Technology* 11, 2 (2009), 105–112. https://doi.org/10.1007/s10676-009-9187-9"

[19] V. Vakkuri, K. Kemell, J. Kultanen, and P. Abrahamsson. 2020. The Current State of Industrial Practice in Artificial Intelligence Ethics. *IEEE Software* 37, 4 (2020), 50–57.

[20] Ville Vakkuri, Kai-Kristian Kemell, and Pekka Abrahamsson. 2019. AI Ethics in Industry: A Research Framework. In *CEUR Workshop Proceedings*. RWTH Aachen University.

[21] Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen, and Pekka Abrahamsson. 2020. "This is Just a Prototype": How Ethics Are Ignored in Software Startup-Like Environments. In *Agile Processes in Software Engineering and Extreme Programming*, Viktoria Stray, Rashina Hoda, Maria Paasivaara, and Philippe Kruchten (Eds.). Springer International Publishing, Cham, 195–210.

[22] Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen, Erika Halme, and Pekka Abrahamsson. 2021. ECCOLA—A method for implementing ethically aligned AI systems. *Journal of Systems and Software* 182 (2021), 111067.

[23] Georg von Krogh. 2018. Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries* (2018).