Preprint notes

Title of the article:
Implementing Ethics in AI: An industrial multiple case study

Authors:
Ville Vakkuri, Kai-Kristian Kemell, Pekka Abrahamsson

Notes:
- This is the author’s version of the work
- The definite version is submitted to 20th International Conference on Product-Focused Software Process Improvement (PROFES 2019) https://profes2019.upc.edu/
Implementing Ethics in AI: An industrial multiple case study

Ville Vakkuri[1] (0000-0002-1550-1110)
Kai-Kristian Kemell[1] (0000-0002-0225-4560)
Pekka Abrahamsson[1] (0000-0002-4360-2226)

1 University of Jyväskylä, PO Box 35, FI-40014 Jyväskylä, Finland
ville.vakkuri|kai-kristian.o.kemell|pekka.abrahamsson@jyu.fi

Abstract. Solutions in artificial intelligence (AI) are becoming increasingly widespread in system development endeavors. As the AI systems affect various stakeholders due to their unique nature, the growing influence of these systems calls for ethical considerations. Academic discussion and practical examples of autonomous system failures have highlighted the need for implementing ethics in software development. However, research on methods and tools for implementing ethics into AI system design and development in practice is still lacking. This paper begins to address this focal problem by providing a baseline for ethics in AI based software development. This is achieved by reporting results from an industrial multiple case study on AI systems development in the health care sector. In the context of this study, ethics were perceived as interplay of transparency, responsibility and accountability, upon which research model is outlined. Through these cases, we explore the current state of practice out on the field in the absence of formal methods and tools for ethically aligned design. Based on our data, we discuss the current state of practice and outline existing good practices, as well as suggest future research directions in the area.

Keywords: Artificial intelligence, AI ethics, AI development, Responsibility, Accountability, Transparency, Developer attitude, Developer awareness.
1 Introduction

The role of ethics in software systems development has dramatically changed following the increasing influence Autonomous Systems (AS) and Artificial Intelligence (AI) systems. AI/AS systems necessitate ethical consideration due to their unique nature. Whereas one can opt out of using conventional software systems, the very idea of being an active user in the context of AI systems is blurred.

The harm potential of these systems, as well as real-life incidents of AI system failures and misuse, have resulted in a growing demand for AI ethics as a part of software engineering endeavors. The growing field of AI ethics has highlighted the need for implementing ethics into AI/AS design. AI ethics studies have argued that AI/AS engineering is not, or, rather, should not be simply a technological or engineering endeavor.[1] The harm potential of AI/AS, and various reported system failures and incidents of misuse have highlighted the need for involving ethics as a part of software engineering in the context of these systems. Specifically, it is argued, developers should be aware of ethics in this context due to their key role in the creation of the system. Aside from discussion among the academia, public voices have also expressed concern towards unethical AI systems in the wake of various real-life incidents (e.g. unfair systems [2]).

Despite the increasing activity on various fronts, a gap between the concerns voiced over AI ethics and SE practice in AI remains in the absence of laws and regulations. It is known that developers are not well-informed on ethics [3]. Developers are known to prefer simple and practical methods, if they utilize any at all [4], which poses a problem in the context of AI ethics where methods and tools for implementing it into practice are lacking. With developers lacking the means to implement ethics into their work, they also lack the means to tackle or even recognize potential ethical issues during development. Not only does this result in potential harm during operations, fixing these issues in production is costly for the developing organization.

To provide empirical data into this on-going discussion on AI ethics, we have conducted a multiple case study on AI system development in the healthcare sector. The goal of this study was to provide a further understanding of the current state of practice, as well as to discover existing good practices that may aid in future method and tool development for AI ethics. The exact research question is formulated as follows: \textbf{RQ: how are AI ethics taken into consideration in software engineering projects when they are not formally considered?}

The rest of this paper is organized as follows: Section 2 describes the background and related work of implementing ethical considerations; Section 3 describes the research framework; Section 4 describes the design and conducted research; Section 5 presents empirical findings; Section 6 discusses the findings in broader context; Section 7 concludes and summarizes the answers for research questions to set the direction for future work.

2 Related work

2.1 The Current State of AI Ethics

Much of the research on AI ethics has been conceptual and theoretical in nature. These studies have e.g. focused on defining AI ethics in a practical manner through various constructs in the form of values. For the time being, this discussion on defining AI ethics has come to center around four values: transparency [5, 6]), accountability [5, 7]), responsibility [5], and fairness (e.g. [2]). Not all four of these values are universally agreed to form the core of AI ethics, however, as we discuss in the following section while presenting our research framework.

Following various real-life incidents out on the field (e.g. Amazon’s biased recruitment AI [8]), AI ethics has also begun to spawn public discussion This has led to governments, standardization institutions, and practitioner organizations reacting by producing their own demands and guidelines for involving ethics into AI development, with many standards and regulations in the works. Countries such as France (Villani Report[9]), Germany (Federal Ministry of Transport and Digital Infrastructure in Germany[10]) and Finland (Finland's age of artificial intelligence[11]) have emphasized the role of ethics in AI /AS, and on an international level the EU began to draft its own AI ethics guidelines which were presented in April 2019 [12]. Moreover, ISO has founded its own ethical, trustworthy AI in ISO/IEC JTC 1/SC 42 Artificial intelligence subcommittee [13]. Finally, some larger practitioner organizations have also presented their own guidelines concerning ethics in AI (e.g. Google guidelines [14]).

Thus far, these various attempts to bring this on-going academic discussion out on the field have been primarily made in the form of guidelines and principles. Out of these guidelines, perhaps the most prominent ones up until now have been the IEEE guidelines for Ethically Aligned Design (EAD), born from the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems alongside its IEEE P7000™ Standards Working Groups, which were branded under the concept of EAD[7].
Past experiences have shown us that guidelines and principles in the field of ICT ethics do not seem to be effective. [3] argued based on empirical data that the ACM ethical guidelines [15] had ultimately had very little impact on developers, who had not changed their ways of working at all. As there is currently little in the way of empirical data on what effects these guidelines may have had on the field, it is possible that this has also been the case with AI ethics guidelines so far. A recent version of the IEEE EAD guidelines [7] already acknowledged this potential gap, noting that a lot of work remained to be done in bridging the gap between research and practice in this area. However, no empirical data regarding the situation out on the field was supplied with this notion.

### 3 Research Model

In addressing ethics as a part of the development of AI and AI-based systems, various principles have been discussed in academic literature. For the time being, the discussion has centered on four constructs: Transparency [5-7], Accountability [5,7], Responsibility [5], and Fairness e.g. [2]. Perhaps notably, a recent EU report [12] also discussed Trustworthiness as its key construct, a value all systems should aim for, according to the report.

The field of AI ethics can be divided into three categories: (1) Ethics by Design (integration of ethical reasoning capabilities as a part of the behaviour of artificial autonomous system, e.g. ethical robots); (2) Ethics in Design (the regulatory and engineering methods supporting ethical implications of AI systems); and (3) Ethics for Design: (codes of conduct, standards, and certification processes that ensure the integrity of developers and users) [16]. In this paper, we focus on the ethically aligned development process, and therefore the latter two categories are focused on in the research framework.

Out of the aforementioned four principles that have been proposed to form the basis of ethical development of AI systems, we consider accountability, responsibility, and transparency (the so-called ART principles) formulated by Dignum [5] a starting point for understanding the involvement of ethics in ICT projects. These three constructs form the basis of ethical AI and attempts to identify their possible relations, as well as relations of other constructs that may be involved in the process.

To make these principles tangible, a subset of constructs in the form of actions (1.1-3.5), discussed in detail in subsection 3.1, were formed under each key concept. These actions were outlined based on the IEEE guidelines for EAD[7]. The actions were split into two categories, Ethics in Design and Ethics for Design(ers), based on Dignum’s [17] typology of AI ethics. Finally, to explain actions we utilized the action and concern components of Abrahamsson’s [17] commitment net model, which is elaborated on in subsection 3.2.

#### 3.1 ART Model

Transparency is defined in the ART principles of Dignum [5] as transparency of the AI systems, algorithms and data used, their provenance and their dynamics. I.e. the transparency refers to understanding how AI systems work by being able to inspect them. Transparency can be argued to currently be the most important of these principles or values in AI ethics. Turilli and Floridi [6] argue that transparency is the key pro-ethical circumstance that makes it possible to implement AI ethics at all. It has also been included into the EAD guidelines as one of the key ethical principles [7].

In the research framework of this study, transparency is considered on two levels: (a) transparency of data and algorithms, as well as, (b) systems development. The former refers to understanding the inner workings of the system in a given situation, while the latter refers to understanding what decisions were made by whom during development.

For the system to be considered transparent (line 1.a), feature traceability (1.1) (EAD’s [7] Principle 5) should be present, and the system should be predictable in its behavior (1.2) (EAD’s [7] Principle 5 & 6). For development to be considered transparent (line 1.c), the decision-making strategies of the endeavor should be clear.
(1.4) [EAD’s [7] Principles 5&6], and decisions should be traceable back to individual developers (1.3) [EAD’s [7] Principle 1, 5, 6]. As a pro-ethical circumstance, transparency also produces the possibility to assess accountability and responsibility (line 1.b) in relation to both development and the system.

Accountability refers to determining who is accountable or liable for the decisions made by the AI. Dignum [5] in their recent works defines accountability to be the explanation and justification of one’s decisions and one’s actions to the relevant stakeholders. In the context of this research framework, accountability is used not only in the context of systems, but also in a more general sense. We consider, for example, how various accountability issues (legal, social) were taken into consideration during the development.

Transparency is required for accountability (line 1.b), as we must understand why the system acts in a certain fashion, as well as who made what decisions during development in order to establish accountability. Whereas accountability can be considered to be externally motivated, closely related but separate construct responsibility is internally motivated. The concept of accountability holds a key role in aiming to prevent misuse of AI and in supporting wellbeing through AI [7].

In our research model, accountability is perceived through the concrete actions of the developers concerning the systems itself. 2.1 Preparing for anything unexpected: (actions that are taken to prevent or control unexpected situation) [EAD’s [7] Principle 8], 2.2 Preparing for misuse/error scenarios (actions that are taken to prevent or control misuse/error scenarios) [EAD’s [7] Principles 7 & 8], 2.3 Error handling (practices to deal with errors in software) [EAD’s [7] Principles 4 & 7]; and 2.4 data security (actions taken to ensure cyber security of system and secure handling of data) [EAD’s [7] Principle 3].

Dignum [5] defines responsibility in their ART principles as a chain of responsibility that links the actions of the systems to all the decisions made by the stakeholders. We consider it to be the least accurately defined part of the ART model, and thus have taken a more comprehensive approach to it in our research framework. According to the EAD guidelines, responsibility can be considered to be an attitude or a moral obligation for acting responsibly [7]. A simplified way of approaching responsibility would be for a developer to ask oneself e.g. “would I be fine with using my own system?”. While accountability relates to the connection between one’s decisions and the stakeholders of the system, responsibility is more focused on the internal processes of the developers not necessarily directly related to any one action. In order to act responsibly, one needs to understand the meanings of their actions. Therefore, in the research framework responsibility is perceived through the actions of the developers concerning, 3.1 perception of responsibility (developers have a sense of responsibility and perception what is responsibility in software development) [EAD’s [7] Principles 2, 4 & 6]; 3.2 distribution of responsibility (who is seen responsible e.g. for any harm caused by the system) [EAD’s [7] Principle 6]; 3.3 encountered problems (how errors and error scenarios are tackled and who is responsible for tackling them) [EAD’s [7] Principles 7, 8]; 3.4 feelings of concern (developers are concerned about issues related to their software) [7]; and 3.5 data sensitivity (developers attitude toward data privacy and data security) [EAD’s [7] Principles 2 & 3].

### 3.2 Commitment

Aspects of commitment were included into the model for data analysis purposes. Specifically, we utilized the commitment net model of Abrahamsson [17] to approach the implementation of ethics into practice and have an explaining theoretical framework to examine ethics role to developers. From this model, we focused on concerns (construct 7) which were analyzed to understand what ethical issues were of interest to the developers. Actions (construct 6) were then studied to understand how these concerns were actually tackled, or whether they were tackled at all.

In commitment net model, actions are connected to concerns because when actions are taken, they are always driven from concerns [17]. On the other hand, however, concerns can exist without any actions taken to address them. This is described by line 8.

The dynamic between actions and concerns was considered a tangible way to approach the focus of this study: practices for implementing AI ethics. Actions were directly likened to practices in this context (lines 6.a & 6.b). On the other hand, concerns were considered of interest in understanding e.g. whether the developers perhaps wanted to implement ethics (lines 7.a & 7.b) but were unable to do so.

### 4 Study Design

This study was carried out as a multiple case study featuring three different cases where AI systems were developed for the needs of the healthcare sector. Descriptions of each case can be found in subsection 1.1 and specifically in Table 1. Data from each case was collected by conducting semi-structured qualitative interviews focusing on
ethics in AI design. The interview questions in their entirety can be found in external resource [18]. We focused on the developer and project point of view by primarily interviewing developers and project managers.

4.1 The Three Cases

Table 1. Case Information

| Case | Case Description                      | Respondent[Reference]         |
|------|--------------------------------------|-------------------------------|
| A    | Statistical tool for detecting marginal-ization | Data analyst [R1]            |
|      |                                      | Consultant [R2]               |
|      |                                      | Project coordinator [R3]      |
| B    | Voice and NLP based tool for diagnostics | Developer [R4]               |
|      |                                      | Developer [R5]               |
|      |                                      | Project manager [R6]         |
| C    | NLP based tool for indoor navigation  | Developer [R7]               |
|      |                                      | Developer [R8]               |

Each case was a specific AI project in a case company. All three projects were development projects focused on creating a prototype of an AI-based healthcare software solution. The AI technologies used were different in each case. In all the tests carried out during the case projects, if any, the users had been informed that the product was a prototype and still under development. Alternatively, any testing had been conducted in a controlled environment.

4.2 Data collection

The interviews were conducted as semi-structured interviews. The interview strategy was prepared using the guidelines of Galletta [19]. The interviews were conducted face-to-face interviews and the audio was recorded. The records were transcribed for the purpose of data analysis. In the transcripts, the cases and respondents were given individual references shown in table 1. We focused on interviewing project personnel, developers and managers, due to our focus on development practices and project issues. The interviews were conducted in Finland, using the Finnish language. This carries with it potential limitations in this context which we touch upon in the discussion section.

4.3 Data analysis

The data from the transcripts were analyzed in two phases. First, we followed a grounded theory (Strauss and Corbin [20] and later Heath [21]) inspired approach. In this phase, the transcripts were coded quote by quote and each quote was given a code describing its contents. The same process was repeated for all eight interviews. In the second phase, we utilized the commitment net model of Abrahamsson [17], discussed in section 2.2, to further analyze and categorize the coded quotes from the first phase.

In using the commitment net model, we focused on the concerns and actions of the developers in relation to software development. Each concern and any actions related to the concern were listed for each respondent. The concerns and actions of each respondent were then compared in order to discover any recurring concerns and actions across respondents and cases.

These findings were then compared to the constructs in our research framework (constructs 1.1 to 3.5) in order to evaluate to what extent AI ethics was being implemented in the project. In this evaluation, actions were emphasized due to the research question of this study. Rather than studying whether the respondents were concerned about the different aspects of AI ethics or not, we focused on their actions by identifying practices, tools, or methods through which they had possibly addressed ethical concerns. I.e. how they had implemented ethics in practice.

Based on this comparison of the research framework subconstructs 1.1-3.5, an evaluation of actions supporting the implementation of AI ethics in the context of these constructs was conducted (Fig. 2). Furthermore, observations related to the research framework were formulated by topic (responsibility, transparency, accountability, and general) in the form of primary empirical conclusions, PECs.
5 Empirical Results

Figure 2 summarizes our results based on a qualitative analysis of the three cases. In this figure depicting our research framework, we have colour coded (black and white) the extent to which the various aspects of AI ethics were taken into account by the case companies. The darkest shade of grey indicates that the area was of concern in all three projects, the second darkest one that it was a concern in two, and the lightest that it was a concern only in one case. In the subsections of this section, we go over our findings in detail through the ART constructs of accountability, responsibility, and transparency.

5.1 Responsibility

The concerns of the developers related to responsibility were varied, but ultimately detached from practice as far as concerns related to AI ethics were considered. The concerns the developers discussed in relation to their idea of responsibility were simply very practical concerns related to internal project matters or delivering a high quality product:

“Responsibility on reporting and keeping the project on schedule” (R6)

PEC1 Developers feel most responsibility towards tackling problems related to software development

On the other hand, as the interviews progressed, the developers did also express some concerns towards various ethical issues. However, these concerns were detached from their current work. They did not affect the way they worked, and the developers felt that these types of concerns were not relevant during development. The presence of concerns in the absence of actions to address those concerns pointed towards a lack of commitment on this front.

“It is just a prototype” (R8)
“I do my best” (R5)
“But this is a prototype, an experiment, just to show people that you can do this type of thing. This doesn’t really have any responsibility issues in it.” (R1)

PEC2 Developers are concerned about the ethical aspects of product development. However, little is done to tackle these concerns.

Furthermore, it was evident that in none of the cases had the hypothetical effects of the system on the stakeholders been discussed. To give a practical example, a system potentially affecting memory illness diagnoses clearly has various effects on its potential users, especially when the test can be taken without supervision. Yet, the developers of this particular tool also felt that their users would not be curious about the workings of the system. They considered it sufficient if the responsibility was outsourced to the user and it was underlined that the system does not make the diagnosis but simply advises doctors.

The developers also exhibited a narrow view of responsibility in relation to harm potential. Only physical harm potential was considered relevant, and the developers felt that none of their systems had such potential.

“Nobody wants to listen to ethics-related technical stuff. No five hour lectures about it. It’s not relevant to the users” (R5)
“I don’t really understand what it [responsibility] has to do with product development. We developers are all responsible.” (R7)
“What could it affect... the distribution of funds in a region, or it could result in a school taking useless action... it does have its own risks, but no one is going to die because of it” (R1)
5.2 Transparency

In terms of transparency of algorithms and data, case A stood out with the team's mathematical knowledge. They utilized algorithms they were familiar with and which they understood on an in-depth level. Thus, the team considered themselves to be able to understand why the system made certain decisions in certain situations. This underlines the importance of mathematical skills in preventing the birth of black boxes in AI development.

“In that sense it’s not really a black box as we can understand what’s going on in there just fine, and we can show the nodes and what affects them. It’s a very transparent algorithm.” (R3)

In the other two cases, the companies utilized existing AI solutions. They did not have an in-depth understanding of the technologies they were utilizing, which resulted in their systems being (partially) black boxes. They understood any components created by the team but did not have a full understanding of the third party components they had used as a base. This presents problems for feature traceability.

PEC4 Black box systems are a typical issue in AI/AS development, which mathematical skills can help avoid

even though transparency of algorithms and data was not present in two of the cases, the developers in case B nonetheless acknowledged its potential importance. However, as it was not considered a formal requirement in the projects, the company did not devote resources towards pursuing it. Even in case A, transparency was not produced as a result of ethical goals but out of business reasons.

“We have talked about the risks of decision-making support systems but it doesn’t really affect what we do” (R5)

PEC5 Developers recognize transparency as a goal, but it is not formally pursued

On the other hand, in relation to transparency of systems development, all three case companies displayed transparency. By having formal decision-making strategies, they were able to keep track of higher-level decisions related to the system. Through proper documentation, they were able to keep track of decisions made on the code level. Version control also assisted in this regard, making it clear who made what changes and when in retrospect. There were thus various existing practices that produced transparency of systems development.

Two of the case companies also acknowledged the effects of team size on transparency of systems development. They noted that, in addition to documentation practices, the small team size itself made it easy to keep track of the actions of individual developers even in an ad hoc manner.

PEC6 Existing SE practices support transparency of systems development

5.3 Accountability

Accountability was actively considered in relation to cybersecurity and data management. The developers were aware that they were in possession of data that was of potential interest to both commercial actors and malicious actors, and that they were accountable for taking measures to keep it secure, as well as to abide to laws related to personal data handling. To this end, cybersecurity was considered as a part of standard company protocol, following established company practices.

“It’s really important how you handle any kind of data, that you preserve it correctly, among researchers, and don’t hand it out to any government actors. For example many of the data packages have kind of interesting data and it can’t get into the wrong hands. I personally can’t see any way to harm anyone with the data we have though” (R2)

“We haven’t really paid much attention to the [data] safety aspects yet… it hasn’t really been a main focus. There’s probably a lot of things we have to take into account [eventually]” (R5)

The respondents were also clearly concerned with error handling, underlined by one of the respondents directly remarking “I aim to make error free software” (R1). The developers were concerned about engineering quality
Software in terms of it being error free, and considered it their professional responsibility to do so. The respondents could highlight various practices they utilized to handle and prevent errors in the project.

**PEC7** Developers feel accountable for error handling and have the means to deal with it

However, error handling was largely considered from the point of view of writing code and testing it in a laboratory setting. I.e. the system was considered error free if there were no red lines in the code in the IDE during development. Only case company B discussed measures they had taken to monitor errors in use. Furthermore, potential misuse (e.g. a prankster drawing a horizontal white line on the pavement to intentionally confuse autonomous vehicles) and error scenarios during the operational life of the system had not been actively considered in any of the case projects.

"The calculations are made in the algorithms so it doesn’t really make mistakes" (R2)

**PEC8** Developers do not take actions to tackle misuse or error scenarios out on the field during development

Due to the nature of machine learning, AI systems learn as they are taught with new data or as they collect it themselves while operating out on the field. From this arises the potential issue of unexpected behavior as a result of machine learning. None of the respondents had made plans to tackle potential unexpected behavior during the operational life of their system, should such behavior arise. In only one of the projects was the possibility directly acknowledged:

"We just put it up for end-users to test and note that this is still being developed" (R7)

**PEC9** Developers do not have plans to deal with unexpected behavior of the system

5.4 Summary of Findings

Past the ART constructs, we highlight some commonalities between the cases on a more general level while summarizing our findings. In none of the cases were ethics implemented by following a formal method or tool, nor were ethical issues considered directly as ethical issues. Rather, any ethical issues tackled in the projects were tackled for practical reasons (e.g. error free software is beneficial from the point of view of customer relations). Nonetheless, some of the ethical issues such as error handling and transparency of systems development were tackled in a systematic manner through existing software engineering practices such as code documentation and version control.

**PEC10** Even when ethics are not particularly considered, development practices have some aspects that support ethically aligned design such as documentation and cybersecurity practices

On the other hand, through ethics were not taken into consideration on a project level, the individual developers exhibited some concern towards socioethical issues arising from their systems. They were able to think of ways their system could negatively affect its users or other stakeholders in its current state. However, they lacked ways of addressing these concerns, as well as ways to conduct such ethical analyses in a systematic and comprehensive manner.

**PEC11** While the developers speculate potential socioethical impacts of the resulting system, they do not have means to address them.

6 Discussion

For the purpose of structuring this section, we have collected all the PECs outlined in preceding analysis section into Table 2 below. We relate each of these findings to existing literature and discuss their implications in this section. We classify each of these PECs based on their contribution into either novel findings, findings that (empirically) validated existing literature, or findings that contradict existing literature.

Many of our findings underline a gap between research and practice in the area. Whereas research on AI ethics alongside various guidelines devised by researchers [7] and practitioners [14] alike has discussed various ethical goals for AI systems, these goals have not been widely adopted out on the field. In this sense, we consider some of our findings (PECs 3, 4, 5, 8, and 9) to contradict existing literature.

For example, extant literature has highlighted the importance of transparency of algorithms and data [5,7,22]. Without understanding how the system works, it is impossible to establish why it malfunctioned in a certain situation, which may e.g. be pivotal in understanding the causes of an accident that resulted in material damage. Our findings point towards transparency being largely ignored as a goal (PEC5). Existing system components are
utilized as black boxes, and developers do not see this as a notable problem (PEC4).

The situation is similar for tackling potential misuse of the systems, error handling during system operations, and handling unexpected system behavior (PEC8-9). These goals are included into the IEEE EAD guidelines [7]. However, none of the case companies took any measures to address these potential issues.

On a further note of transparency, however, the lack of emphasis placed on it is also curious in relation to feature traceability in SE. For decades, understanding the inner workings of the system was considered key in any SE endeavor [23]. Yet, in the context of AI systems, the long-standing goal of feature traceability seems to be waning. Our findings point towards this being at least partially a result of a lack of mathematical understanding, as the one case company that considered their system to be fully transparent also noted that they fully understood the mathematics behind the algorithms they utilized. In using existing components in their systems, developers may not always understand the algorithms in these components. Indeed, in this vein, [24] noted that simply seeing the code is not enough if the algorithm is not understood, or the system is not understood as a whole.

Table 2. List of Primary Empirical Conclusions (PECs)

| PEC | Theoretical component | Description | Contribution |
|-----|-----------------------|-------------|--------------|
| 1   | Responsibility        | Developers feel most responsibility towards tackling problems related to software development | Empirically validates existing literature |
| 2   | Responsibility        | Developers are concerned about the ethical aspects of product development. However, little is done to tackle these concerns. | Novel |
| 3   | Responsibility        | Responsibility of developers and development is under-discussed. | Novel |
| 4   | Transparency          | Black box systems are a typical issue in AI/AS development, which mathematical skills can help avoid | Contradicts existing literature |
| 5   | Transparency          | Developers recognize transparency as a goal, but it is not formally pursued | Contradicts existing literature |
| 6   | Transparency          | Existing SE practices support transparency of systems development | Empirically validates existing literature |
| 7   | Accountability        | Developers feel accountable for error handling and have the means to deal with it | Empirically validates existing literature |
| 8   | Accountability        | Developers do not take actions to tackle misuse or error scenarios out on the field during development | Contradicts existing literature |
| 9   | Accountability        | Developers do not have plans to deal with unexpected behavior of the system | Contradicts existing literature |

This is the author’s version of the work. The definite version is submitted to PROFES 2019 for peer review
Even when ethics are not particularly considered, development practices have some aspects that support ethically aligned design such as documentation and cybersecurity practices.

While the developers speculate potential socioethical impacts of the resulting system, they do not have means to address them.

Though we discovered various examples of ethics not being implemented, we also discovered that various existing and established SE practices can be used to implement AI ethics (PEC10). Documentation, version control, and project management practices such as meeting transcripts produce transparency of systems development by tracking actions and decision-making (PEC6). Similarly, software quality practices help in error handling also in the context of AI ethics (PEC7), although they do not specifically account for the errors autonomous systems may face while operating out on the field.

While discussing responsibility with the respondents, we also discovered that most of their responsibility was related to producing quality software and meeting project requirements. This validates existing literature in the area of SPI (e.g. Unterkalmsteiner, [25]). Notably, we also discovered that the developers had ethical concerns towards their systems, which is a novel finding in this context (PEC2). Little is currently known about the state of practice out on the field, although a recent version of the EAD guidelines speculated about a gap in the area, which our findings support in relation to most aspects of AI ethics. Despite AI ethics largely not being implemented, our findings point towards it partially being a result of a lack of formal methods and tools to implement it (PEC11).

Thus, following this study, as well as a past case study [4], we suggest that future research seek to tackle the lack of methods and tooling in the area. Though developers may be concerned about ethical issues, they lack the means to address these concerns. On the other hand, methods can also raise the awareness of developers in relation to AI ethics, creating concerns where there now are none. In creating these methods, we suggest exploring existing practices that can be used as is or tailored to implement AI ethics, as we have discussed here (PEC10).

Given the amount of activity in AI ethics currently, with many governmental actors drafting their own AI ethics guidelines, likely followed by regulations, methods and tools will likely have practical demand in the future. Thus, even if one barrier to implementing AI ethics is currently the fact that it is seldom considered a requirement on a project level, regulations and laws can force organizations to take ethics into account. This would inevitably result in a demand for methods in this area, as well as the birth of various in-house ones.

There are some limitations that should be mentioned. Importantly, the outlined research model and its subset of constructs are heavily based on ART principles of Dignum [5] and IEEE’s EAD [7]. This effects view taken on AI ethics and may exclude some parts of ethical considerations related to ethical AI development. However, the EAD with its versatile group of authors can be seen as a distilled version of the ongoing AI ethics discussion considering most relevant parts of the discussion.

7 Conclusions & Future work

In this paper, we have sought to better understand the current state of practice in AI ethics. Specifically, we studied the way AI ethics are implemented, if at all, when they are not formally considered in a software engineering project. To this end, we conducted a multiple case study featuring three case companies developing AI solutions for the healthcare sector.

Based on our data, we discovered that some existing good practices do exist for some aspects of AI ethics. For example, current practices out on the field are already capable of producing transparency of systems development. Moreover, the developers are aware of the potential importance of ethics and exhibit concerns towards ethical issues. Yet, developers lack the tools to address these concerns. As tackling ethics is not a formal requirement in AI projects, these concerns largely go unaddressed for business reasons as well.

In this light, we consider the creation of methods and tools for implementing AI ethics important. These will both help developers implement AI ethics in practice as well as raise their awareness of ethical issues by e.g. helping them understand harm potential outside the narrow scope of physical harm.
References

1. Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., Loh, J., Winfield, A. F. T., Yampolskiy, R.: Towards moral autonomous systems, arXiv Preprint arXiv:1703.04741, 2017.

2. Flores, A. W., Bechtel, K., Lowenkamp, C. T.: False positives, false negatives, and false analyses: a rejoinder to "Machine bias: there's software used across the country to predict future criminals, and it's biased against blacks", Federal Probation 80(2), 38 (2016).

3. McNamara, A., Smith, J., Murphy-Hill, E.: Does ACM's code of ethics change ethical decision making in software development? Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering ESEC/FSE pp. 729-733, 2018. doi:10.1145/3236024.3264833

4. Vakkuri, V., Kemell, K., Kultanen, J., Siponen, M., Abrahamsson, P.: Ethically Aligned Design of Autonomous Systems: Industry viewpoint and an empirical study. (2019) Preprint arXiv:1906.07946

5. Dingnum, V.: Responsible Autonomy. (2017). Preprint arXiv:1706.02513.

6. Turilli, M., Floridi, L.: The ethics of information transparency. Ethics and Information Technology, 11(2), 105-112 (2009) doi:10.1007/s10676-009-9187-9.

7. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE. (2019). https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html, last accessed 2019/6/14.

8. Reuters: Amazon scraps secret AI recruiting tool that showed bias against women. (2019). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scrap-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, last accessed 2019/6/14.

9. Villani, C., et al. For a meaningful artificial intelligence: towards a French and European strategy. Conseil national du numérique, (2018). https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf, last accessed 2019/6/14.

10. German Federal Ministry of Transport and Digital Infrastructure: Automated and Connected Driving. (2017). https://www.bmvi.de/DE/Topics/Digital-Matters/Automated-Connected-Driving/automated-and-connected-driving.html, last accessed 2019/5/13.

11. Finland's age of artificial intelligence Report (2017). https://www.tekoalyaika.fi/en/reports/finlands-age-of-artificial-intelligence/, last accessed 2019/6/14.

12. AI HLEG (High-Level Expert Group on Artificial Intelligence): Ethics guidelines for trustworthy AI. (2019) https://ec.europa.eu/eigital-single-market/en/news/ethics-guidelines-trustworthy-ai, last accessed 2019/6/14.

13. ISO/IEC JTC 1/SC 42 Artificial intelligence, webpage, https://www.iso.org/committee/6794475.html last accessed 2019/6/14.

14. Pichai, S. AI at Google: our principles. Blog (2018). https://www.blog.google/technology/ai/ai-principles/ 2019/5/13.

15. Goterbarn D. W., Brinkman B., Flick C., Kirkpatrick M. S., Miller K., Vazansky K. & Wolf M. J. ACM Code of Ethics and Professional Conduct. (2018) https://www.acm.org/code-of-ethics. 2019/6/14., 2019/6/14. last accessed 2019/5/25.

16. Dingnum, V.: Ethics Inf Technol 20:1. (2018) doi:10.1007/s10676-018-9450-z

17. Abrahamsson, P.: Commitment Nets in Software Process Improvement. Annals of Software Engineering, 14, 407-438. (2002). doi:10.1023/A:1020526329708.

18. Developer questionnaire, External resource webpage, http://users.jyu.fi/~vimavakk/AIDevQuestionnaire, last accessed 2019/6/14.

19. Galletta, A.: Mastering the semi-structured interview and beyond: From research design to analysis and publication. NYU press. (2013).

20. Strauss, A., Corbin, J.: Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, 2nd Ed. Sage Publications Inc. Thousand Oaks, CA, US, (1998).

21. Heath, H.: Developing a grounded theory approach: a comparison of Glaser and Strauss, Int. J. Nurs. Stud., 41(2), 141-150, (2004).

22. Bryson, J., Winfield, A.: Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems, Computer, 50(5), 116-119, (2017)

23. Gotel O. et al. Traceability Fundamentals. In: Cleland-Huang J., Gotel O., Zisman A. (eds) Software and Systems Traceability. Springer, London (2012).

24. Ananny, M., Crawford, K.: Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." New Media & Society 20(3) 973–89. (2018): doi:10.1177/1461444816676645.

25. Unterkalmsteiner, M. et al.: Evaluation and measurement of software process improvement—a systematic literature review. IEEE Transactions on Software Engineering, 38(2), 398-424 (2011).