Ethically Aligned Design of Autonomous Systems: Industry Viewpoint and an Empirical Study

Ville Vakkuri
Kai-Kristian Kemell
Joni Kultanen
Mikko Siponen
Pekka Abrahamsson

Abstract

Progress in the field of Artificial Intelligence (AI) has been accelerating rapidly in the past two decades. Various autonomous systems from purely digital ones to autonomous vehicles are being developed and deployed out on the field. As these systems exert a growing impact on society, ethics in relation to artificial intelligence and autonomous systems have recently seen growing attention among academia. However, the current literature on the topic has focused largely on theoretical contributions, and there is a gap between research and practice in the area. Though this gap has been acknowledged in existing studies, the exact issues resulting in this gap remain blurred. In order to better understand the gap in the area, we conduct a multiple case study of five case companies. Based on the data, we highlight a number of issues in the area in terms of implementing AI ethics in practice. We then propose ways to tackle this gap.

Key Words: Ethics, artificial intelligence, autonomous systems, software development, companies, guidelines

Introduction

Artificial Intelligence (AI) systems and Autonomous Systems (AS) are becoming increasingly ubiquitous. Most inhabitants of the developed world interact with AI systems on a daily basis. The more sophisticated recommendation systems utilized by various B2C Software-as-a-Service media platforms such as YouTube utilize AI and Machine Learning (ML), and specifically Deep Learning (DL), to generate personalized recommendations for their users. Autonomous Vehicles (AVs) operated by AI are slowly entering the public roads, AI-based surveillance systems armed with facial recognition capabilities are already being deployed, and various AI systems are being invested in and developed across fields such as medicine (Zhang et al., 2022). In general, progress in AI has been far faster than anticipated by experts in the past.

One key difference between AI/AS and conventional software systems is that the idea of an active user is often blurred. One seldom uses AI systems as opposed to being an object to their data collection procedures or other actions. Whereas one can opt out of using conventional software systems, one often has little control over being targeted by AI systems. Moreover, some AI systems are Cyber-Physical Systems (CPS) that operate both in the digital and physical world. CPSs are various, ranging from security cameras to cargo ships, and exhibit various degrees of autonomy. CPSs such as AVs are now entering public spaces where they can interact with passers-by and cause physical damage rather than being confined to e.g., factories as factory robots (Charisi et al., 2017).

Given their potentially enormous societal impact, AI systems should be designed while taking ethics into consideration (Bostrom & Yudkowsky, 2018; Bryson & Winfield, 2017; The IEEE Global Initiative, 2019). For example, when an AV gets into an accident, we should always be able to understand why. This is not always simple even with full access to the program code as ML systems can be highly complex even to their creators (Ananny & Crawford, 2018). Another factor that makes ethical consideration challenging at times is that the effects of the systems are not always direct (e.g., effects of individual AV on its surroundings vs. societal effects caused by 50% of the traffic being AVs).

Awareness of AI ethics issues has recently been growing in the wake of various practical incidents. For example, YLE, the Finnish national public broadcasting company, commissioned and deployed an AI-based moderation system to replace its human moderators for user comments. It was not until the system was deployed in practice and started making decisions that issues began to manifest to the point where the system was rather quickly decommissioned. This is but one of many incidents where an AI system is designed, developed, and deployed, only for it to prove unusable due to issues related to AI ethics. Similarly, users are becoming more aware of data privacy issues and are more conscious of what their data is being used for and whom it is being collected by.

As a result of the growing interest towards AI ethics related issues, a large number of guidelines have been devised to help organizations tackle AI ethics issues. These guidelines have been developed by companies, the academia, and governments (Jobin et al., 2019). IEEE's Ethically Aligned Design (EAD) (The IEEE Global Initiative, 2019) is among these guidelines, and has been developed as a part of a particularly extensive initiative. As methods in the area remain highly technical, focusing on only subsets of the development process (Morley et al., 2020), these guidelines have become the primary tools for implementing AI ethics for the time being.

However, though both academic and public discussion in the area of AI ethics has accelerated, the state of practice in the area remains unclear. In a past study, we argued that a gap between research and practice in the area exists, based on quantitative survey data (Vakkuri et al., 2020). In this paper, we take a closer look at this gap to better understand the issues companies face in implementing AI ethics. Specifically, we study the current industry mindset in relation to AI ethics

from the point of view of some of the most common AI ethics principles discussed in AI ethics guidelines, including IEEE's EAD (The IEEE Global Initiative, 2019). The exact research question of this paper is formulated as follows:

RQ: What practices, tools, or methods, if any, do industry professionals utilize to implement ethics into AI design and development?

The rest of this article is structured as follows. In the next section, we discuss the theoretical background of the study. In the third section, we discuss the research design. In the fourth section, we present our results, the implications of which we then discuss in the fifth section. The sixth and final section concludes the paper.

Background

In this section, we discuss the context of this study. In the first subsection, we discuss the current state of AI ethics. In the second subsection, we discuss AI in the context of Autonomous Vehicles (AVs). In the third and final subsection, we discuss commitment, which was used as the research framework for data analysis in this study.

The Current State of Ethics in AI and Ethically Aligned Design

The ethics of AI is a long-standing area of ethical discussion in ICT ethics. This discussion has accelerated notably in the past decade following technological progress in the area. As AI systems become increasingly sophisticated, hypothetical AI ethics scenarios of the past are becoming practical issues.

Indeed, researchers from various disciplines have voiced concerns over ethics in AI systems (Borenstein et al., 2021). Following various incidents out on the field, public voices of concern have also been heard. The general public is, for example, becoming increasingly aware of data privacy issues and the way their data is handled by companies. The General Data Protection Regulation (GDPR), while not AI-specific, does end up affecting AI systems among others given how reliant most current AI systems are on large masses of data.

Laws and regulations, however, do generally tend to be slow in the face of technological progress. Some companies have already begun to consider AI ethics, publishing their own AI ethics guidelines or statements online (many of which were reviewed by Jobin et al. (2019)). It remains largely unknown to what extent these guidelines are then really employed in practice inside these organizations, but some companies are at least aware of some of the current AI ethics issues. Aside from companies (e.g., Google (Pichai, 2018)), governments (e.g. EU (AI HLEG, 2019)), and standardization institutions have also begun to work on and publish guidelines intended to help organizations implement AI ethics in practice. One such notable initiative has been the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, which has since branded the concept of Ethically Aligned Design (EAD) and published a set of guidelines (The IEEE Global Initiative, 2019) featuring various principles for AI ethics, distilling much of the recent academic discussion into another set of guidelines.

These guidelines have been initial attempts at creating tools to help organizations implement AI ethics in practice. As much of the academic research on AI ethics has been conceptual and theoretical, focusing on defining and structuring AI ethics through principles and values, bringing this discussion to industry organizations presents evident challenges. The various guidelines published so far have summarized this discussion into principles, although these principles can still be difficult for developers to implement in practice. Indeed, the entire idea behind using these guidelines to implement AI ethics has been criticized (Mittelstadt 2019).

On the other hand, methods in the area ethical AI/ML are largely technical, focused mainly on managing machine learning and other subsets of the development process (Morley et al., 2020). Though this is important, such methods do not help with the big picture in developing ethical AI systems. In the absence of AI ethics methods to direct the development process as a whole, the aforementioned AI ethics guidelines such as EAD (The IEEE Global Initiative, 2019) have become common as tools for implementing AI ethics. Numerous such guidelines exist, and though they discuss different principles, some consensus in the area already exists. (Canca, 2020; Jobin et al., 2019).

Indeed, the ongoing academic discussion on ethics in AI has so far converged on different principles, some of which are also discussed in EAD (The IEEE Global Initiative, 2019). Jobin et al. (2019), based on their analysis of 84 AI ethics guidelines, argued that the following principles were the most common ones, in a descending order of popularity: (1) transparency, (2) justice, fairness and equality, (3) non-maleficence, (4) responsibility and accountability, (5) privacy, (6) beneficence, (7) freedom and autonomy, (8) trust, (9) sustainability, (10) dignity, and (11) solidarity. In our analysis, we utilize transparency, accountability, and responsibility, as well as what we argue can be considered a subset of transparency, predictability, as a framework for the data collection in this study (as we discuss again in the third section).

Transparency is the central AI ethical construct present in most AI ethics guidelines (Jobin et al., 2019). Turilli and Floridi (2009) argue that it is, in fact, the pro-ethical circumstance that makes it possible to implement AI ethics in the first place. Very related to transparency is also the idea of explainable AI systems, which has recently been discussed extensively both in academia and among practitioners (e.g. Adadi & Berrada, 2018; Rudin, 2019).

We consider there to be two types of transparency: (1) transparency of algorithms and data (Dignum, 2017) (i.e., the transparency of systems), and (2) transparency of systems development (i.e., decision-making etc.). Predictability can be considered a subset of transparency, as the EAD guidelines do (The IEEE Global Initiative, 2019), and as we thus do in our analysis. As the word implies, it refers to whether the system acts predictably. For example, if an autonomous coffee machine successfully brews coffee 8 times out of 10, we are left wondering what happened the other two times and why.

Accountability and responsibility are in some ways related, though still separate constructs. Accountability focuses on who is accountable or liable for the decisions made by the AI. Dignum (2017), in her work, defines accountability to be the explanation and justification of one's decisions and one's actions to the relevant stakeholders. Transparency is required for accountability, 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, responsibility is internally motivated. Responsibility can be considered to be an attitude or a moral obligation for acting responsibly (The IEEE Global Initiative, 2019). In order to act responsibly, one has to weigh their options and consciously evaluate the effects of their actions and decisions.

These three main constructs (Transparency, Accountability,

and Responsibility) and one sub construct (Predictability) are our focus in this study. They are AI ethics principles that have become some of the most prominent ones commonly featured in the numerous AI ethics guidelines currently in existence (Jobin et al., 2019). We discuss this choice further in the research design section that follows.

To conclude this section, we further position this paper in this area. While ethics in AI has become a prominent topic among the academia, as well as in public discussion, the current state of industrial practice remains unclear. In another study, we argued that there is a gap between research and practice in the area (Vakkuri et al., 2020). However, the exact nature of this gap is not clear. The focus of this paper is to further explore the situation in the industry and to begin tackling the present lack of tooling for EAD and other AI ethics guidelines. By better understanding the gap in the area, we are able to provide better tools to tackle the issues out on the field.

Artificial Intelligence and Autonomous Vehicles

Currently, AVs are being developed across industries. Though arguably the most media exposure is on cars given their nature as B2C personal vehicles, the possibilities of AI have been explored in relation to drones, cargo ships, buses, trains, and airplanes alike. While the degree of autonomy exhibited by various types of vehicles is steadily increasing, fully autonomous vehicles are still rarely used in practice. Such vehicles are actively being tested in various fields, however.

Safety in these systems is a justified and widely acknowledged concern (Nascimento et al., 2020). Regardless of software quality in AVs, accidents and dangerous situations are inevitable. Such situations may, for example, result from faulty sensors. However, whereas human actors seldom have time to make a carefully thought-out decision in the face of an impending accident, and may sometimes be too slow to properly react at all, AI systems are capable of making a decision near instantaneously. Thus, such systems are required to make difficult ethical decisions in situations where accidents are inevitable one way or the other (Evans et al., 2020) This includes dilemmas such as the commonly cited example "Should Your Car Kill You to Save Others?" (Bonnefon, 2016; Lo Piano, 2020).

From the point of view of AI ethics, the AI ethics principles discussed in the preceding subsection also apply in the context of AVs. Accountability, for example, can be argued to be even more relevant when material damage is a possibility. Similarly, data and data-related issues are also relevant for AVs.

In practice, ethical issues are ultimately left for the developers to tackle. Though company level policies and guidelines can direct development work, micro-level decisions are nonetheless left to individual developers. Thus, developers working with AI need to be able to implement ethics into the systems they develop. This calls for both awareness of AI ethics among developers, as well as tools to implement it (Vakkuri et al., 2021). Currently, little is known about how AI ethics is handled in practice in organizations.

Commitment

As the theoretical framework for this study, we approach ethics in AI through the lens of commitment. In industrial psychology and organizational behavior, commitment is a long-standing area of research (Benkhoff, 1997). The idea of commitment has been of interest primarily because of the assumption that the commitment of employees relates to performance. O'Reilly and Chatman (1986) remark that "although the term commitment is broadly used to refer to antecedents and consequences, as well as the process of becoming attached and the state of attachment itself, it is the psychological attachment that seems to be the construct of common interest." Drawing from this, we consider commitment to be the attachment an individual feels towards an object (organization, ideal etc.).

Aside from behavioral studies from fields such as psychology, commitment has been studied in the past in relation to software process improvement (SPI) (Abrahamsson, 2002). Abrahamsson (2002) proposed a model of commitment nets (Figure 1, p. 6). The model suggests that drivers, both internal and external, may result in concerns which would then manifest as actions, and those actions would then lead to both intended and potentially unintended outcomes. Commitment, in this model, can be observed when concerns result in actions. We utilize this commitment net model as the theoretical framework of this study, as we discuss in detail in the next section.

Research Design and Protocol

This study was carried out as a multiple case study of five cases (Table 1, p. 7). Each case company develops AI systems, although

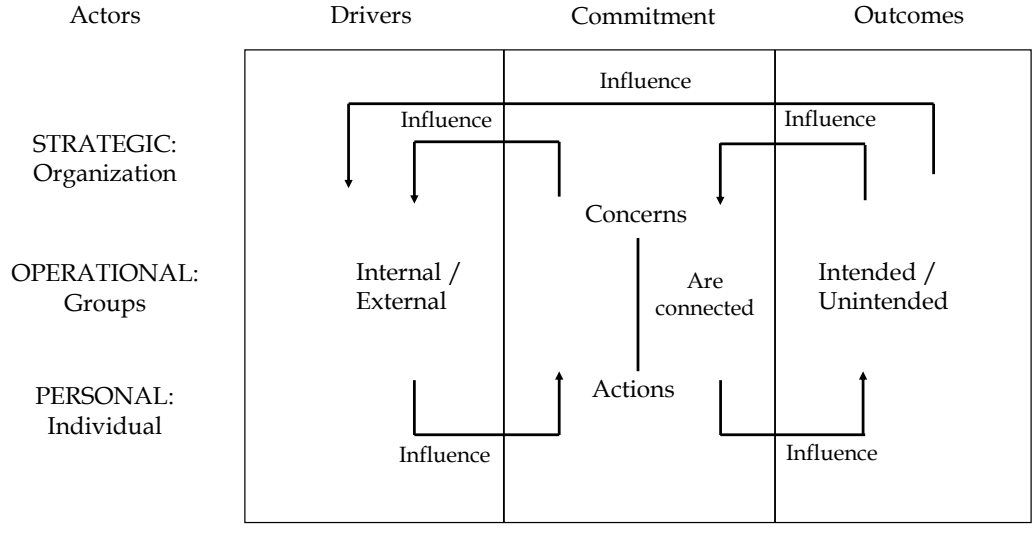

Figure 1. The Commitment Net Model of Abrahamsson (2002).

in different fields, or only as a portion of their operations. Data were collected via semi-structured qualitative interviews. The interview instrument in its entirety can be found in the Appendix 1 (p. 15).

In short, the interview protocol (Appendix 1, p. 15) was designed to focus on the key constructs discussed in background section: transparency, accountability, responsibility, and predictability. We avoided directly discussing ethics as different individuals have different conceptions of what ethics is in this context. This is underlined by the on-going academic discussion as well (see for example (Friedman et al., 2013). Instead, we focused on asking practical questions related to these ethical principles.

In devising this interview instrument, we chose to focus on some of the earlier AI ethics themes that have remained prominent. The principles we focused on in the interview instrument were common in the various guidelines reviewed by Jobin et al. (2019) and are present in IEEE's EAD (2019) as well. Thus, though we focus on only some AI ethics principles, the principles utilized here are some of the most central ones. We discuss this research framework behind the interviews in detail in another paper (Vakkuri et al., 2019).

#	Company Description	Respondent [Reference]
1	Large, >400 employees; Software, Generic	Data Scientist [R1]; Senior Data Scientist [R2]
2	SME (Small/Micro), <25 employees; Software, Healthcare	Development Lead [R3]
3	SME (Small/Micro), <25 employees; Software, Process Industry	CTO [R4]
4	Large (Multinational), >100 000 employees; Consulting	Functional Designer [R5]
5	Large (Multinational), >25000 employees; Vehicle Industry	AI Development Lead [R6]

Table 1. Case Company Information.

We utilized the commitment net model of Abrahamsson (2002) (see Section Commitment) as the theoretical framework for the analysis of these cases. We approached commitment through the concerns that the employees might have had towards implementing ethics in AI design, as well as through the actions they might have taken as a result of their concerns.

To analyze the data, we used the grounded theory method inspired by (Corbin et al., 2014). After the interview, data was coded to classify themes and we focused on concerns the respondents had towards AI ethics issues. Then, after identifying concerns, we looked at what actions the respondents (or their organizations) had taken to tackle these concerns - if any at all. By doing so, we sought to understand whether any commitment towards ethics in AI design existed in the case companies. To give a practical example, if one indicates concern towards losing weight, but exhibits no actions such as making dietary changes or exercising, there is no commitment present.

However, the goal of this analysis was not to find out whether the organizations exhibited any commitment towards AI ethics. Rather, we focused on the actions the respondents and their organizations had taken to address their concerns. In this fashion, we wanted to identify any practices, tools, or methods that had been used to address ethical concerns, i.e., to find out how the respondents had implemented AI ethics.

In our analysis of the data, we summarize our findings through what we refer to as Primary Empirical Contributions (PEC). We consider these to be findings that are worth noting despite occasionally being outside the direct scope of our research question. These PECs are then further discussed in the discussion section and provide a framework for it.

Empirical Results

Our interviews of the case companies indicated that the industry is aware of the potential importance of AI ethics. Every respondent agreed that ethics is useful. However, the case companies had highly differing views on how relevant it was in practice, and none of them remarked using development practices that directly supported implementing it. This underlined that the companies did not have clear tools or methods for implementing ethics. This disconnect seemed, in part, to also stem from a lack of consensus on what (AI) ethics actually referred to. As a part of our empirical results, we elaborate some of our findings with relevant quotes from the respondents. However, our findings are not solely based on the quotes, but on our data in general.

> …I actually try to use the word 'ethics' as little as possible because it's the kind of word that everyone understands in their own way, and so they can feel that it's not relevant to what we're doing at all… [R4]

> …the discussion on AI ethics doesn't really affect most … excluding maybe Google and some others like that … the AI really isn't at the level where it would really necessitate in-depth ethical consideration [R3]

PEC1: Ethics is considered important in principle, but as a construct it is considered detached from the current issues of the field by developers. In other words, the on-going academic discussion on AI ethics has not reached the industry at large.

Only the respondent involved in developing a medical AI system had a more practical view of ethics in relation to their current project. However, the respondent noted that the ethical consideration had already been carried out externally. Indeed, fields such as the field of medicine inherently have very strict regulations regarding, for example, data management, leaving little leeway for developers to make their own ethical decisions:

> We have in-house quality measurements and these regulation requirements are very strict, so these things pretty much come as a given for us. And, of course, if you think about it the other way, we consequently think about these things [ethics] even less because we already have such clear regulations and requirements for what we do [R3]

PEC2: Regulations force developers to take into account ethical issues while also raising their awareness of them.

On the other hand, though ethics as a construct was considered impractical and too theoretical, the respondents did all nonetheless concern themselves with various constructs related to AI ethics (in this case: transparency, predictability, accountability, and responsibility). These constructs were considered practical by the respondents, as we discuss in the following subsections.

Transparency

All case companies were concerned with both transparency of systems and transparency of systems development. Furthermore, transparency of systems was considered both from the point of view of developers and users. However, the actions taken to address these concerns (if any) were varied (Table 2, p. 8) across cases:

> The most important thing is that we can see directly how it works, and that it's trackable, now, and later. [R5]

> …it is typically a little un-transparent how the decisions are made. Of course we can analyze them, but due to the complexity of the neural network architecture, it's a little difficult to accurately explain why it did something. [R6]

Whereas transparency from the point of view of developers was considered in relation to e.g., the algorithms and the neural network architecture, transparency from the point of view of the users was considered on a less technical level (Table 3, p. 8). The respondents felt that the users had little reason to be able to see inside the system or the so-called black box as such. It was considered more important that the users would be able to understand how it works on the very basic level:

> Our systems are aimed at these… operational personnel, like the paper plant guys down on the factory floor […] they don't really know what happens inside the system and we don't feel that they really need to know, either […] they just understand that, okay now all this data goes in, and the suggestions are then based on that data [R4]

> …the users won't really notice a difference compared to the earlier systems they have used. We just want to offer them better and more timely data. So that's of course one question: how to make it clear for them that there are some uncertainties there so that they don't expect the information to always be perfect. But… I don't really know how much of a problem this is -- I haven't really spoken to our end-users [R5]

PEC3: Developers have a perception that the end-users are not tech-savvy enough to gain anything out of technical system details.

In terms of transparency of systems development, four of the five companies indicated clear concern towards it and had taken actions to address the concern (Table 4, p. 9). Largely, (code) documentation was considered to be the primary way of producing transparency in the development process by making it apparent who made what changes, why, and when. Additionally, conducting audits was discussed as one tangible practice for producing transparency in the development process. This was one of the few areas where a consensus among the companies could be observed in ethical practices.

PEC4: Documentation and audits are established Software Engineering project practices that form the basis in producing transparency in AI/AS projects.

Predictability

One of the main concerns shared by all respondents was the potential unpredictability of the system (Table 5, p. 9). The respondents discussed clear actions they had taken to either avoid unpredictable behavior, to mitigate it, or to prevent it in the future in case it takes place. An example of such an action can be ML management by means of using different sets of training data or by limiting its utilization.

Driver	Actor	Concern	Action(s)
Project need	R1	Keeping the system understandable to developers (i.e. transparency to developers)	No recognized actions
Legislation; Regulations	R3		Devoting time to understanding the training data
Company need	R4		Devoting time to understanding the AI used as a template for the system; Building analytics into the system
Company need	R5		No recognized actions; (Planned future action: documentation)
Company need	R6		Devoting time to understanding the training/testing data; Mode verification

Table 2. Commitment Towards Transparency to Developers.

Driver	Actor	Concern	Action(s)
Project need	R1	Keeping the system understandable to the end users (i.e. transparency to users)	No recognized actions
No clear driver	R2		Educating the customer/user
Market edge; Process improvement	R3		No recognized actions
Company need; Professionalism	R4		Educating the customer/user
Company need	R5		Writing helpful system descriptions
Company need; Professionalism	R6		Educating the customer/user; Communication with customer/user

Table 3. Commitment Towards Transparency to Users.

> …we have even cut some functionalities […] of the system in order to make it more predictable, which has reduced the amount of unexplained results we have gotten out of it […] in practice we've been able to explain all of the faulty results so far [R3]

PEC5: Machine learning is considered to inevitably result in some degree of unpredictability. Developers need to explicitly acknowledge and accept heightened odds of unpredictability.

When discussing steps taken to avoid unpredictability, the respondents also discussed their concerns related to a hypothetical situation in which the system has already acted unpredictably (Table 6, p. 10). All six respondents and five case companies had outlined some courses of action for such a scenario, although some of the actions pointed towards a lack of commitment (e.g., apologizing and reacting on a case-by-case basis is a very ad hoc plan).

Finally, in relation to predictability, four of the respondents discussed cyber security threats as a part of unpredictable system occurrences (Table 7, p. 10), even if they are caused by external actors as opposed to the system itself. Indeed, in the case of especially CPSs, cybersecurity threats can pose life-threatening danger if e.g., an autonomous bus is hijacked digitally. Given that cybersecurity is a longstanding area of research and industry practice, companies generally have established policies and even cybersecurity departments for dealing with cybersecurity issues. Thus, few actionable measures or practices were underlined by the respondents in response to their actions in tackling cybersecurity concerns.

Accountability and Responsibility

The consensus among the respondents was that no system could be completely fault-free, with five respondents expressing concern towards potential harm caused by their system(s) (Table 8, p. 11). Most respondents could also list some actions their organization had taken to either avoid or mitigate harm caused by their system. However, some of the respondents felt that their system(s) had no direct potential for harm even if it did act unpredictably or wrongfully, due to it e.g., being a purely digital business intelligence system.

PEC6: Developers consider the harm potential of a system primarily in terms of physical harm. Potential systemic effects are often ignored.

Additionally, the respondent working on healthcare AI (R3) indicated a more personal approach to responsibility than the other respondents as they felt that they were directly responsible for the well-being of some of their users.

PEC7: Physical harm potential motivates personal drivers for responsibility.

Notably, the respondents ultimately outsourced the responsibility and/or accountability to their users despite exhibiting a commitment to mitigate or prevent harm. They felt that they had taken what measures they could to prevent harm, and that it was then up to the user to stay safe (e.g., doctors should be critical of the suggestions of medical AI):

PEC8: Main responsibility is outsourced to the user, regardless of the degree of responsibility exhibited by the developer.

As the respondents discussed having concerned themselves and their project teams very little with direct discussions about

Driver	Actor	Concern	Action(s)
Project need; Customer need	R1	Keeping track of who does and decides what and why (I.e., transparency of development)	Documentation
Project need; Customer need	R2		Documentation; Conducting audits; Distinct roles in development team
Customer need; Market need; Regulations	R3		Documentation; Conducting audits, audit trail
Company need	R5		Documentation
Company need	R6		Launch of new management process

Table 4. Commitment Towards Transparency of Development.

Driver	Actor	Concern	Action(s)
No clear driver	R1	System acts unpredictably (i.e., preventing an incident)	Awareness of unpredictability; Recognizing what errors are acceptable; Preparedness for incidents of unpredictability
Company need	R2		Representative training data; Training for designer
No clear driver	R3		Reduce functionalities and complexity of system; Narrow the scope of use of machine learning
No clear driver	R4		Accept the (minimal) odds of unpredictability; Acknowledging that statistical tools also make mistakes; Root cause analysis
No clear driver	R5		Using the system only in confined spaces
Company need	R6		AI/ML model validation

Table 5. Commitment Towards Preventing Unpredictability.

ethical matters related to their systems, they did not consider responsibility strongly from an ethical point of view. Instead, they approached responsibility largely from the point of view of delivering a product that fulfilled expectations set for it (Table 9, p. 11) internally, by various stakeholders, or by regulations. Some of the respondents also felt that delivering a quality product was their responsibility as professionals of the field.

PEC9: Developers typically approach responsibility pragmatically from a financial, customer relations, or legislative point of view rather than an ethical one.

Discussion

We have collected the Primary Empirical Contributions (PECs) outlined in the results section into Table 10 (p. 12). They have been split into three categories based on their contribution: (1) empirically validates existing literature, (2) contradicts existing literature, and (3) new knowledge. Overall, the primary contribution of this study is its empirical approach focusing on developers and the state of practice. Existing studies in the area have been largely theoretical.

The most general finding of this study is that it further confirms that there is a gap between research and practice in the field of AI ethics (PEC1). The academic discussion on AI ethics and the values related to it (transparency, etc.) seems to not have affected the industry yet. This is consistent with the findings of McNamara et al. (2018) who concludes that the ACM Code of Ethics (Gotterbarn et al., 2018) has done little to change the way developers work. Whittlestone et al. (2019), Mittelstadt (2019) and Canca (2020), also argue that guidelines are likely to be difficult to implement in practice out in the field. Moreover, we have also argued that there is indeed such a gap in the area in another paper with a quantitative approach (Vakkuri et al. 2020). There thus seems to be a clear gap between research and practice in the area. The rest of the findings of this study serve to further our understanding of said gap.

We argue that this gap largely stems from a lack of tooling and methodologies in the area, as has been suggested by Whittlestone et al. (2019) as well. Based on our data, industry professionals currently address ethical issues through various ad hoc practices. While numerous guidelines exist (Jobin et al., 2019), they are not actionable (Whittlestone et al., 2019; Canca 2020) and consequently see little use. Tools and methods are needed to make them actionable. Currently, tools and methods in the area offer little help in designing ethical AI systems and managing the big picture, as they focus on the technical aspects of the development such as managing ML (Morley et al., 2020).

To help in tackling this gap in practice, we have begun to work on a method to help implement AI ethics in practice. This meth-

Driver	Actor	Concern	Action(s)
Customer need; Company need	R1	System makes mistake in production (i.e. hypothetical scenario in which an incident took place)	Accept the (minimal) odds of unpredictability; Be willing to react; Apologize
Company need; Project need; Professionalism	R2		Be willing to react; Apologize; [Planned future action: communication/ action plan]
Customer need; Financial	R3		Feedback options to product development; Using mistake as example in learning data; Accept the unlikely unpredictability; Acknowledging that statistical tools also make mistakes
No clear driver	R4		Piloting before full release; Reacting feedback and fixing issues; Narrowing functionalities in design
Company need; Customer need	R5		Piloting oversite; Cutting system functionalities; Fixing bugs when noticed
Company need; Customer need; Legislation	R6		Backup systems

Table 6. Commitment Towards Addressing an Incident of Unpredictability.

Driver	Actor	Concern	Action(s)
Company need; Customer need	R1	Cybersecurity / Data security / Adversary attacks	Follow quality process and corporate policy
Company need; Project need; Professionalism	R2		Recommendations on how to prepare; Awareness of context of use (i.e., who can do what with the system)
Company need; Customer need; Legislation	R3		Follow quality process and corporate policy
Company need; Customer need	R6		Backup systems; Preparing for attacks

Table 7. Commitment Towards Cybersecurity.

od, ECCOLA, that builds on existing research, has been developed by researchers and applied in industry projects. We have published this method in another paper (Vakkuri et al., 2021). It is an on-going initiative, and though ECCOLA is still being developed further, it has reached a state of maturity where we wish to share the method with the scientific community, as well as the industry.

Aside from tooling, one way of addressing this gap would be through changes in legislation and regulations (PEC2). However, legislative changes are slow and may struggle to keep up with the advances in technology. They may also have negative, limiting effects on AI development (e.g., regulations on international waters limit testing maritime AVs). Nonetheless, legislation and regulations are starting to address AI issues, with the General Data Protection Regulation (GDPR) and the upcoming AI Act affecting AI systems in the EU area.

However, it should nonetheless be noted that some companies do seem to utilize these AI ethics guidelines. Nagadivya et al. (2020) studied companies using AI ethics guidelines to guide AI system development and argue that they can be useful in doing so. Arguably, the guidelines certainly do provide a starting point for implementing AI ethics, even if it takes effort from the organization to make them actionable. It would seem, though, that most organizations currently do not wish to devote resources towards doing so.

Indeed, based on our findings, it seems that developers currently do not approach ethics in a systematic manner and do not utilize any tools or methodologies to implement it. However, ethical values discussed in academic literature are nonetheless taken into account in the industry to some extent. According to the IEEE EAD guidelines (The IEEE Global Initiative, 2019), documentation is a key in producing transparency. This was also acknowledged by all case companies (PEC4), although the sufficiency of their documentation remains unknown. Similarly, the challenges ML poses to system predictability are discussed in existing literature and also acknowledged by industry professionals (PEC5).

On the other hand, while the IEEE EAD guidelines (The IEEE Global Initiative, 2019) and other such guidelines typically encourage transparency in terms of providing users with technical details of the systems as well, developers feel that their users do not possess the technical knowledge to make any use of said information (PEC3). Here the opinions of the developers also notably contradict existing literature in which transparency has been extensively discussed e.g., from the point of view of the users or the general public being able to understand the technical side of the system.

In terms of responsibility, developers do not seem to possess the skills to evaluate the harm potential of AI systems comprehensively. They exhibit a narrow view of the harm potential of such systems, focusing on physical harm (PEC6). This is a topic that has not been extensively studied thus far but practical incidents do point towards this being the case. In other words, either developers are unaware of these issues or they are simply ignored, e.g., in favor of financial gain. While developers exhibit more responsibility if they consider the system to have physical harm potential (PEC7), social and emotional impacts of AI systems are ignored (PEC6). Developers also do not consider the systemic effects of AI systems, which can be important (German Federal Ministry of Transport and Digital Infrastructure, 2017). This further highlights the gap in the area, as AI ethics literature discusses the harm potential of AI systems extensively and takes into account social issues such as racial bias (See for FAccT community focusing fairness, accountability, and transparency in socio-technical systems).

However, we do feel that one cannot expect developers to conduct such comprehensive ethical analysis unassisted and without training. Training developers (or university students

Driver	Actor	Concern	Action(s)
Customer need	R1	Responsibility for potential harm caused by the system or a specific algorithm	Adhering to contracts; Responsible project management
Company need; Project need; Personal	R2		No recognized actions
Personal	R3		Accept the (small) odds of harm; Communication with the customer to minimize the risk of harm
No clear driver	R5		Design the system so that even wrong decisions are not harmful
No clear driver	R6		Minimizing potential harm; Accept small odds of harm; Build a system that produces less harm than humans in the same context

Table 8. Commitment Towards Responsibility for Potential Harm.

Driver	Actor	Concern	Action(s)
Company need; Commercial; Professionalism	R1	Delivering a working product / Delivering what was promised	Setting realistic goals for the system
Commercial	R3		No recognized actions
Company need; Customer need; Professionalism	R4		Piloting; Keeping the human in the loop
No clear driver	R5		Discussion inside project team; Communication with customer

Table 9. Commitment Towards Addressing an Incident of Unpredictability.

#	Theoretical component	Description	Contribution
1	Conceptual	Ethics is considered important in principle, but as a construct it is considered detached from the current issues of the field by developers.	Empirically validates existing literature
2	Conceptual	Regulations force developers to take into account ethical issues while also raising their awareness of them.	Empirically validates existing literature
3	Transparency	Developers have a perception that the end-users are not tech-savvy enough to gain anything out of technical system details.	Contradicts existing literature
4	Transparency	Documentation and audits are established Software Engineering project practices that form the basis in producing transparency in AI/AS projects.	Empirically validates existing literature
5	Transparency	Machine learning is considered to inevitably result in some degree of unpredictability. Developers need to explicitly acknowledge and accept heightened odds of unpredictability.	Empirically validates existing
6	Responsibility; Accountability	Developers consider the harm potential of a system primarily in terms of physical harm. Potential systemic effects are often ignored.	New knowledge
7	Responsibility; Accountability	Physical harm potential motivates personal drivers for responsibility.	Empirically validates existing literature
8	Responsibility; Accountability	Main responsibility is outsourced to the user, regardless of the degree of responsibility exhibited by the developer.	New knowledge
9	Responsibility; Accountability	Developers typically approach responsibility pragmatically from a financial, customer relations, or legislative point of view rather than an ethical one.	New knowledge

Table 10. Primary Empirical Conclusions of the Study.

who will go on to become developers in the future) to take into account AI ethics and teaching them how to do so is important. Additionally, carrying out such ethical analyses calls for distribution of work in organizations, or even hiring ethical experts to carry out the analysis (Canca 2020). Furthermore, we once more underline the importance of tools and methods in this regard.

Moreover, in relation to responsibility, developers seldom consider responsibility important purely for ethical reasons. Rather than being concerned about being ethical, they are concerned about potential financial losses or bad publicity resulting from the system being unethical (PEC9). This is to some extent similar to how companies have approached environmental issues or business ethics at large, although nonetheless new in the specific context of AI ethics. Companies are more likely to tackle these issues for financial or legislative reasons, as opposed to doing so simply to act responsibly. This should be considered when attempting to raise awareness of AI ethics in the industry.

Regardless of the degree of responsibility exhibited by the developers, the responsibility is ultimately outsourced to the user(s) of the system (PEC8). In other words, the developers feel that the user should always be critical towards the suggestions of the system, whether the user is a doctor or a factory worker, and that how they use the system is their responsibility. Similar lines of argumentation are seen, for example, in relation to firearm legislation, and thus while this is new in the context of AI ethics, outsourcing responsibility in this sense as a phenomenon is not novel.

Also, outsourcing responsibility in this context is interesting when combined with PEC3, as the developers simultaneously feel that their end-users are not tech savvy enough to benefit from being explained or shown the technical details of the system. Yet, despite the users thus having no in-depth understanding of how the systems work, the developers feel that the users should be able to evaluate the actions of the systems in an informed fashion. This issue has been, in part, acknowledged in existing literature. Scholars have repeatedly voiced their concerns over black boxes and demanded explainable AI systems. (Bryson & Winfield, 2017; Adadi & Berrada, 2018). Recently the demand has even switched beyond explainable AI and ML models to interpretable models (Rudin, 2019).

In terms of future research directions, we recommend any studies seeking to address the evident gap between research and practice in the area. This includes further studies into the state of practice (e.g., further studies on how companies implement AI ethics when using AI ethics guidelines to do so), as well as tools or methods for implementing AI ethics.

Limitations of the Study

The generalizability of the findings is always an issue for qualitative case studies. Given the qualitative approach of this study, we cannot claim that our results would be representative of the current state of the industry at large with 5 case companies involved. However, we would turn to Eisenhardt (1989) who argues that for novel research areas, five cases is an acceptable number.

Empirical studies in AI ethics, including those looking into the current state of the art, are currently still few in number and there seems to be a gap in the area between research and practice (see for example (Vakkuri et al., 2020) or (Morley et al., 2020), which leads us to argue that this is a novel area of research.

Another limitation is, still related to these case companies, that all the case companies were either Finnish or international companies whose Finnish branch was the only one involved in this study. This is a potentially notable limitation in this context because much of the discussion on AI ethics has been US-based. Therefore, it is possible that especially US companies

might be more concerned with AI ethics than companies based in Finland. However, in another study (Vakkuri et al., 2020), we have taken on a quantitative approach to studying the current state of practice and did not find any notable differences between Finnish and US companies.

Finally, the research framework used in this study presents some limitations as well. In particular, the construct of ethics can impose threat to the validity of this study as ethics and values have tendency to mean different things to different individuals (Friedman et al., 2013). In an attempt to tackle this limitation, the concept of ethics was approached through more context related sub-constructs (grounded in existing research) and questions directly mentioning ethics were kept to a minimum. As much of the research so far in AI ethics has focused on defining principles for ethical AI systems, existing research in the area offered various concepts that could be used for this purpose. In this study, we have utilized, but some of these (transparency, accountability, responsibility, and predictability). While these themes are central, with e.g., transparency being the most high-profile one (Jobin et al., 2019), there are various other principles associated with AI ethics. Our approach, thus, only focused on some aspects of AI ethics. Additionally, while planning the interview protocol and conducting the data collection, we have mostly kept our distance as researchers, maintaining a distinct role and doing our best to only collect data while avoiding advising or leading the participants on into any direction.

Conclusions

In this paper, we have conducted a case study to understand the current state of practice in relation to ethics in AI. The case study featured five case companies, in which the data was gathered through semi-structured, qualitative interviews. We utilized the commitment net mode and grounded theory to analyze the data through the concerns the organizations or individuals exhibited towards various ethical issues, as well as the actions they had taken to address said concerns.

In summary, developers consider ethics important in principle. However, they consider ethics as a construct impractical and distant from the issues they face in their work. There is thus a clear gap between research and practice in the area as the developers are not aware of the academic discourse on the ethics of AI.

The key finding of this study was that none of the case companies utilized any tools or methodologies to implement AI ethics. Based on our data, it seems that developers lack ways to systematically implement AI ethics into practice. They tackle ethical issues separately from other development tasks and in an ad hoc fashion, using highly differing practices across organizations. While various guidelines for AI ethics currently exist, written by both practitioners and scholars alike, these guidelines are not used by industry experts. One reason behind this lack of adoption is likely the fact that these guidelines consist of principles and values rather than actionable practices, which can make them challenging to utilize in practice. At very least, this results in a situation where organizations hoping to utilize these guidelines in practice must devote resources towards first making them actionable for the developers.

We recommend that future studies seek ways to make these guidelines, or AI ethics in general, actionable for the industry. This could be achieved in a number of ways. For example, methods and tools can help organizations implement AI ethics in practice. Alternatively, among other options, a maturity model for AI ethics focusing on processes could also help in this regard. Ultimately, and in any case, it seems that guidelines may not be the way to proceed and that we should look elsewhere when it comes to making AI ethics practical. A large number of AI ethics guidelines already exists and it is unlikely that any new set of guidelines would provide a notable contribution at this point.

References

Abrahamsson, P. (2002) "Commitment nets in software process improvement", Annals of Software Engineering, Vol. 14, No. 1, pp. 407-438.

Adadi, A. and Berrada, M. (2018) "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)", IEEE Access, Vol. 6, pp. 52 138-52 160.

AI HLEG (High-Level Expert Group on Artificial Intelligence) (2019) "Ethics guidelines for trustworthy ai". Available https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

Ananny, M. and Crawford, K. (2018) "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability", New Media & Society, Vol. 20, No. 3, pp. 973-989

Balasubramaniam, N. and Kauppinen, M. and Kujala, S. and Hiekkanen, K. (2020) "Ethical guidelines for solving ethical issues and developing ai systems", Product-Focused Software Process Improvement, pp. 331-346.

Benkhoff, B. (1997) "Disentangling organizational commitment: The dangers of the ocq for research and policy", Personnel Review, Vol. 26, No. 1, pp. 114-131.

Bonnefon, J. F. and Shariff, A., and Rahwan, I. (2016). "The social dilemma of autonomous vehicles. Science", Vol. 352(6293), pp. 1573-1576.

Borenstein, J., Grodzinsky, F.S., Howard, A., Miller, K.W., & Wolf, M.J. (2021). "AI ethics: A long history and a recent burst of attention", Computer, 54(1), pp. 96-102.

Bostrom, N. and Yudkowsky, E. (2014), "The ethics of artificial intelligence", in Frankish, K. and Ramsey, W.M. (Eds.), The Cambridge handbook of artificial intelligence, Cambridge University Press. pp. 316-334.

Bryson, J. and Winfield, A.F. (2017) "Standardizing ethical design for artificial intelligence and autonomous systems", Computer, Vol. 50, No. 5, pp. 116-119.

Canca, C. (2020). "Operationalizing AI ethics principles", Communications of the ACM, 63(12), pp. 18-21.

Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M. Sombetzki, J. and Winfield, A.F. and Yampolskiy, R. (2017). Towards moral autonomous systems. arXiv preprint. Available https://doi.org/10.48550/arXiv.1703.04741

Corbin, J. and Strauss, A. (2014). "Basics of qualitative research: Techniques and procedures for developing grounded theory, Sage publications.

Dignum, V. (2017) "Responsible autonomy", arXiv preprint. Available https://doi.org/10.48550/arXiv.1706.02513

Eisenhardt, K. M. (1989) "Building theories from case study research", The Academy of Management Review, Vol. 14, No. 4, pp. 532-550.

Evans, K. and de Moura, N. and Chauvier, S. and Chatila, R. and Dogan, E. (2020) "Ethical decision making in autonomous vehicles: The av ethics project", Science and Engineering Ethics.

Friedman, B. and Kahn, P. H. and Borning, A., and Huldtgren,

A. (2013), "Value sensitive design and information systems", in Doorn, N. and Schuurbiers, D. and Van de Poel, I. and Gorman, M. E. (Eds.), Early engagement and new technologies: Opening up the laboratory, Springer, Dordrecht. pp. 55-95.

German Federal Ministry of Transport and Digital Infrastructure (2017). "Automated and Connected Driving". Available https://www.bmvi.de/EN/Topics/Digital-Matters/Automated-Connected-Driving/automated-and-connected-driving.html

Gotterbarn, D. W. and Brinkman, B. and Flick, C. Kirkpatrick, M. S. and Miller, K. and Vazansky, K. and Wolf, M. J. (2018) "Acm code of ethics and professional conduct", Association for Computing Machinery. Available https://www.acm.org/code-of-ethics

Jobin, A. and Ienca, M. and Vayena, E. (2019) "The global landscape of ai ethics guidelines", Nature Machine Intelligence, Vol. 1, No. 9, pp. 389-399, 2019.

Lo Piano, S. (2020) "Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward", Humanities and Social Sciences Communications, Vol. 7, No. 9.

McNamara, A. and Smith, J. and Murphy-Hill, E. (2018) "Does ACM's code of ethics change ethical decision making in software development?" Proceedings of the 2018 26th ACM ESEC/FSE, pp. 729-733.

Mittelstadt, B. (2019) "Principles alone cannot guarantee ethical ai", Nature Machine Intelligence, pp. 1-7.

Morley, J. and Floridi, L. and Kinsey, L. and Elhalal, A. (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, Vol.26 No.4, pp. 2141-2168.

Nascimento, A. M. and Vismari, L. F. and Molina, C. B. S. T. and Cugnasca, P. S. and Camargo, J. B. and d. Almeida, J. R. and Inam, R. and Fersman, E. and Marquezini, M. V. and Hata, A. Y. (2020) "A systematic literature review about the impact of artificial intelligence on autonomous vehicle safety", IEEE Transactions on Intelligent Transportation Systems, Vol. 21, No. 12, pp. 4928-4946.

O'Reilly, C. A. and Chatman, J. (1986) "Organizational commitment and psychological attachment: The effects of compliance, identification, and internalization on prosocial behavior", Journal of Applied Psychology, Vol. 71, No. 3, pp. 492-499.

Pichai, S. (2018) "Ai at google: our principles". Available https://www.blog.google/technology/ai/ai-principles/

Rudin, C. (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", Nature Machine Intelligence, Vol. 1(5), pp. 206-215.

The IEEE Global Initiative, (2019) "Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, first edition", Available https://standards.ieee.org/content/ieee-standards/en/industryconnections/ec/autonomous-systems.html

Turilli, M. and Floridi, L. (2009) "The ethics of information transparency", Ethics and Information Technology, Vol. 11, No. 2, pp. 105-112.

Vakkuri, V. and Kemell, K-K and Kultanen, J and Abrahamsson, P (2020) "The current state of industrial practice in artificial intelligence ethics", IEEE Software, Vol. 37, No. 4, pp. 50-57.

Vakkuri, V. and Kemell, K-K. and Abrahamsson, P. (2019) "Ai ethics in industry: research framework", CEUR Workshop Proceedings. RWTH Aachen University, Vol. 2505. Available http://ceur-ws.org/Vol-2505/paper06.pdf

Vakkuri, V. and Kemell, K-K. and Jantunen, M. and Halme, E. and Abrahamsson, P. (2021). "ECCOLA — A method for implementing ethically aligned AI systems", Journal of Systems and Software, Vol. 182, 111067.

Whittlestone, J. and Nyrup, R. Alexandrova, A. and Cave, S. (2019) "The role and limits of principles in ai ethics: Towards a focus on tensions", Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 195-200.

Zhang, D., Maslej, N., Brynjolfsson, E., Etchemendy, J., Lyons, Terah., Manyika, J. Ngo, H., Niebles J.C., Sellitto, M., Sakhaee, E., Shoham, Y., Clark, J., and Perrault, R. (2022) "The AI Index 2022 Annual Report", AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University Available https://aiindex.stanford.edu/report/

APPENDIX 1 – AI Developer Questionnaire

General

1. What kind of software does your organization develop?
2. To whom are they developed to? / Who uses them? (customers / in-house projects)
3. How is AI involved in the software development? (AI / AI based solutions?)
4. What is your own role in the development?

Accountability

5. How much can you personally affect the functionalities of the AI solutions and the decisions made on them?
6. Who makes the final decisions concerning the development? (Such as what functionalities are good and what to choose to use?)
7. If the AI solution causes harm or damage to the user or third parties, who is responsible?
   a) How much responsibility do you consider to be on you, based on your role in the organization
8. Are there other questions or issues on accountability that you have considered within your organization in relation to the development process or the end-products?

Predictability

9. How well do you consider the behavior of your AI solutions can be predicted beforehand? Could there be or has there been unexpected behavior to be noticed?
10. How do you prepare for this kind of unexpected behavior or possible malfunctions, and how do you react to them if they occur?
11. What is the level of acceptable risk or damage in case of malfunctions to the end-users or third parties?
12. How have you considered possible cases of misuse or abuse of your product? What could they be?

Transparency

13. How well the development process is being documented? For instance, can certain functions or decisions made during the development process be led back to the individuals behind them?
14. Are all the actions made by the AI solution transparent in a sense, that the logic behind the functions can be understood? (For example, the algorithms used and how they perform the reasoning – also during exceptions in functionalities.)
15. How well do the end-users know what the AI solution does and how it does it?

AI Ethics

16. Has your organization already faced some ethical issues or questions regarding AI development, and what have they been?
17. Do your organizational policies consider ethical aspects within AI development, and how?
18. How does the consideration of ethical aspects show in practice in the development process?
19. Do you consider taking ethical aspects into account in AI development would be beneficial to your organization? How?

Authors

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

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

Joni Kultanen, University of Jyväskylä, joni.m.kultanen@jyu.fi.

Mikko Siponen, University of Jyväskylä, mikko.t.siponen@jyu.fi.

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