

How Do AI Ethics Principles Work? From Process to Product Point of View

Kai-Kristian Kemell^{1,*}, Ville Vakkuri² and Fahad Sohrab³

¹University of Helsinki, Finland

²University of Vaasa, Finland

³Tampere University, Finland

Abstract

Discussing the potential negative impacts of AI systems and how to address them has been the core idea of AI ethics more recently. Based on this discussion, various principles summarizing and categorizing ethical issues have been proposed. To bring these principles into practice, it has been common to repackage them into guidelines for AI ethics. The impact of these guidelines seems to remain small, however, and is considered to be a result of a lack of interest in them. To remedy this issue, other ways of implementing these principles have also been proposed. In this paper, we wish to motivate more discussion on the role of the product in AI ethics. While the lack of adoption of these guidelines and their principles is an issue, we argue that there are also issues with the principles themselves. The principles overlap and conflict and commonly include discussion on issues that seem distant from practice. Given the lack of empirical studies in AI ethics, we wish to motivate further empirical studies by highlighting current gaps in the research area.

Keywords

AI ethics, Machine Learning, Principles, Product

1. Introduction

In the past decade, AI systems have become increasingly ubiquitous across highly varied use contexts. As these systems exert more and more influence in society, it also becomes increasingly important to minimize any negative impacts they might also have. Minimizing the negative impact of AI systems is particularly vital in the case of systems that impact the general public, such as autonomous vehicles and medical systems.

Discussing the potential negative impacts of AI systems and how to address them has been the core idea of AI ethics more recently. In the past, AI ethics was primarily focused on discussing future scenarios, but following technological progress, these future scenarios are increasingly becoming a reality alongside other issues. Numerous conceptual AI ethics studies have highlighted potential issues in different types of AI systems, and discussed ways to begin addressing them. This discussion has resulted in a number of AI ethics principles [1]. However,


Conference on Technology Ethics – Tethics, October 18–19, 2023, Turku, Finland

*Corresponding author.

✉ kai-kristian.kemell@helsinki.fi (K. Kemell); ville.vakkuri@uwasa.fi (V. Vakkuri); fahad.sohrab@tuni.fi (F. Sohrab)

🆔 0000-0002-0225-4560 (K. Kemell); 0000-0002-1550-1110 (V. Vakkuri); 0000-0002-8080-4011 (F. Sohrab)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

📄  CEUR Workshop Proceedings (CEUR-WS.org)

applying ethical principles into practice is a recurring challenge in computer ethics in general [2] [3]), and is arguably one in AI ethics as well [4] [5] [6].

To bring AI ethics into practice, guidelines have been the primary way of approaching AI ethics. Numerous guidelines have been built around the AI ethics principles proposed in conceptual papers (and by companies and other organizations) [1]. The impact of these guidelines remains small, however [6] [4] [3] [5]. In addition to guidelines, some tools have been developed to help with the technical implementation of principles [6]. However, these are individual, precise technical tools [6] rather development approaches. For example, recent technical tools include "tactics for fair clustering and producing intersectionally fair rankings, as well as testing the probabilistic fairness of pre-trained logistic classifiers, assessing "leave-one-out unfairness" or measuring robustness bias among many others." [7] Some Software Engineering (SE) methods to implement AI ethics have also recently been proposed, including the RE4AI Ethical Guide [8] and ECCOLA method [9].

The practice remains a key issue in AI ethics. Only a small number of empirical studies on AI Ethics has been published. These studies have primarily focused on the current state of practice in AI ethics by looking at AI development practices (e.g., [4] [10] [11]). The two SE methods we are aware of [9] [8] have also been developed using empirical data. Studies looking at the successful implementation of AI ethics, in particular, are lacking [7].

Even though empirical studies in AI ethics in general remain scarce, we feel that is nonetheless time to broaden our focus from the development process to the product itself. While any further studies on developing ethical AI are certainly still sorely needed, we feel that discussing the role of the product is relevant at this stage. As some guidelines, methods, and other artifacts that (are claimed to) help develop ethical AI exist – and are hopefully also being applied in practice – we should look at their impact not only on the development process but on the product or service itself.

Yet we are not aware of any studies looking at the impacts of these ethical tools, guidelines, and methods on operational AI/ML systems that have been developed using them. Even if the development process is changed to support the development of ethical AI, by, e.g., using AI ethics guidelines (which in and of itself is a recurring challenge [12]), what kind of impact would this have on the resulting system? If we were to take the numerous AI ethics principles currently considered important and implement them into a system to make it ethical, what would the system look like?

In this paper, we discuss what we consider a gap in AI ethics research: the point of view of the product. Thus far, much emphasis has been placed on the *development* and design of ethical AI, with guidelines seeking to direct high-level design and development decisions – although their practical relevance has been questioned [5] [4] [12] – and individual, precise technical tools [6]. In comparison, little attention has been paid to the finished systems and their operational life. Consequently, little is known about whether these principles really even *could* be implemented by the book. Measuring adherence to these principles is similarly a challenge to be tackled, although not one we can begin to address in this paper. To illustrate our points regarding principles, we look at a number of high-profile AI ethics principles and evaluate, through a hypothetical AI/ML system, what they could mean in practice if a system were to be developed by fully implementing them into practice. Based on this thought exercise, we discuss potential issues within these guidelines and principles and propose future avenues

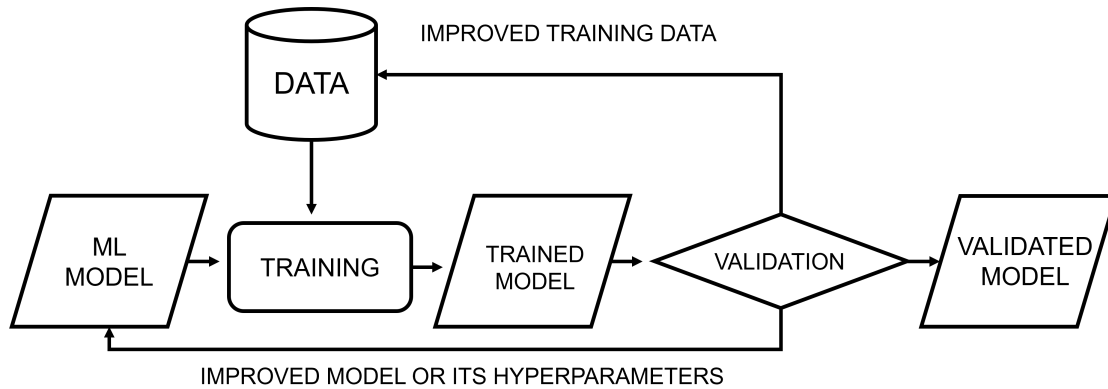


Figure 1: ML training process according to Mikkonen et al. [13]

of research in AI ethics.

2. Background: AI system development

When discussing AI systems in this paper, we refer to Machine Learning (ML) systems specifically. To illustrate the process of creating ML models, we turn to Mikkonen et al. [13], who present one way of conceptualizing the process in their paper. This process is depicted in Figure 1. In order to train ML model(s), training data is needed to optimize the model(s) parameters. This data needs to be bought, collected, or otherwise acquired. Depending on how the data was acquired, it may need to be cleaned and otherwise processed before training the model. During training, the model(s) learn the unknown function that maps the input data to the output values. Much depends on the ML approach being used, as, for example, supervised learning where one has "access to the data and to the "right answer" often called a label, e.g., a photo and the objects in the photo" [13] requires highly refined data. In contrast, in unsupervised learning, the unknown mapping function is identified without leveraging the labels of data. The aims of the system also need to already be clear during the training stage so that the right type of data can be decided on.

The training process is iterative, as ML models are iteratively trained, adjusted, and validated until they are able to produce satisfactory results in terms of, e.g., prediction accuracy or any other metric(s) of choice. Validation can be carried out with different data.

This process is at the core of any ML model development. However, as DevOps (a portmanteau of Development and Operations), and continuous SE in general, continue to be state-of-the-art in SE, the idea of continuous SE is also prevalent in ML development in the form of MLOps [14] (a portmanteau of ML and Operations). Consequently, various papers discussing ML development also discuss operations in the context of ML and involve issues related to (continuous) deployment and monitoring of ML systems. In their paper seeking to better define MLOps John et al. [15] highlight three pipelines for MLOps: Data Pipeline, Modeling Pipeline, and Release Pipeline, each of which also includes a number of sub-processes. Compared to the process seen

in 1, for example, this conceptual model includes the third release pipeline that deals with the deployment and operational life of the models in addition to their creation.

These processes are a part of the software development process in general. ML components are only one part of the entire system, while much of the work is conventional SE [16]. In practice, ML is often seen as a separate process that takes place alongside the conventional software development process. There are thus two processes running in parallel: (1) ML model building and (2) software development. Simplifying to some extent, ML models are built and evaluated by the data science team, while the rest of the development is carried out by the development (and operations) team(s). This is also highlighted by Lwakatere et al. [17] as they discuss the relationship between ML and SE: "increasingly there is a need to integrate the development workflow of ML system into existing SE processes and methods, such as agile methods and continuous integration (CI) practice."

As we argue for a more product-focused point of view on AI ethics in this paper, we should also briefly discuss what 'product' means in this context. As we have established in this regard, we discuss ML systems specifically when speaking of AI systems. Additionally, when speaking of products in the context of ML systems, we specifically refer to systems that are operational. As opposed to looking at requirements or design decisions, which has often been the point of view in AI ethics, we want to look at the finished systems and how they are or should be affected by AI ethics. We, therefore, refer to ML systems in the deployment and/or monitoring stages.

3. AI ethics: from guidelines to requirements and practice

Having discussed ML development in general in the preceding section, we discuss AI ethics in this system. In the first subsection, we provide an introduction to AI ethics principles. In the second subsection, we discuss the practical implementation of these principles.

3.1. AI ethics principles: what should an ethical AI look like?

Currently, AI ethics is commonly approached through various principles. These AI ethics principles act as a way of categorizing AI ethics issues and are commonly distilled into guidelines containing multiple principles each. Such guidelines are produced by companies, researchers, as well as (supra)national actors such as the EU [18][1][6]. Jobin et al.[1] and Hagendorff[18] have conducted reviews of a large number of these guidelines in their papers.

Various principles have guided discussions on AI ethics and continue to do so. Despite a large number of principles that have been proposed over time, the discussion on AI ethics principles has recently begun to converge on a set of recurring principles whose definitions are also starting to unify to some extent [6]. In Table 1, we list the most common AI ethics principles based on three widely cited AI ethics papers discussing AI ethics principles.

The first paper in Table 1 is the AI ethics guidelines review of Jobin et al.[1], which provides an extensive review and summary of the most commonly utilized principles. The authors also offer synthesized definitions for each of the principles included in the paper. The second paper is the guideline review of Hagendorff [18]. This review also looks at a large number of guidelines but is more focused on simply quantifying the principles in terms of how frequently they appear

Table 1

Ethical principles identified in existing AI guidelines

Jobin et al. [1]	Hagendorff [18]	Morley et al.[6]
Transparency		Explicability
Justice and fairness	fairness	Justice
Non-maleficence		Non-Maleficence
Responsibility accountability	Accountability	-
Privacy	Privacy	-
Beneficence	-	Beneficence
Freedom and autonomy	-	Autonomy
Trust	-	-
Sustainability	-	-
Dignity	-	-
Solidarity	-	-
-	Robustness or safety	-

in guidelines, while also providing some meta-analysis of the authors of the guidelines. The third paper is that of Morley et al.[6] where the authors compare AI ethics guidelines and methods (here 'method' refers to highly specific ML methods as opposed to SE methods such as SCRUM), and based on the results, propose their own requirements for ethical AI in the form of principles.

Out of these three papers, we utilize the paper of Jobin et al. [1] as the framework for our principle review in the following section. This paper is utilized due to its extensiveness and due to the extensive descriptions of each principle proposed by the authors. These definitions provide us a way of utilizing these principles without having to focus on arguing about their definitions in this paper, as they are based on synthesis in a scientific publication. We focus on exploring the implementation of the principles as opposed to defining them.

3.2. Implementing AI ethics in practice through principles

AI ethics research aims to facilitate the creation of ethical AI systems ultimately. For the time being, principles have been the main way of working towards this goal. By distilling principles into guidelines, principles have been intended to serve as a way of implementing AI ethics in practice.

However, implementing principles directly into practice is challenging, even with the help of guidelines [2] [5] [6] [4]. Indeed, perhaps the largest problem typically associated with guidelines and principles in AI ethics, or computer ethics in general, is that it is difficult to make principles actionable for developers. In more detail, Mittelstadt [5] argues that there are four challenges associated with using principles in AI ethics: "(1) common aims and fiduciary duties, (2) professional history and norms, (3) proven methods to translate principles into practice, and (4) robust legal and professional accountability mechanisms." From the point of view of SE research, the second and third challenges are the most relevant ones out of these four.

The first challenge is related to the fact that software organizations have no obligation to do good and are, in fact, more responsible to their board and shareholders than the general

public [5]. Principles and guidelines are only useful if there is an intent (or obligation) to use them. Aside from legal obligations to do so, incentives for implementing AI ethics are largely personal moral convictions or fear of reputational risk. The lack of legal obligations to implement ethics is the fourth challenge. For companies, it is often 'enough' to simply adhere to laws and regulations, and consequently, these laws and regulations (e.g., GDPR) are important in setting the bar for the bare minimum in ethics [19].

As for the second challenge, professional history and norms, the challenge is that there is far less tradition for being a 'good' developer, for example, than there is for being a 'good' doctor [5]. Some attempts to establish such a tradition have been made, e.g., in the form of the ACM Code of Ethics [20]. Much like the first challenge, this is primarily related to (the lack of) intent to implement AI ethics, particularly from the point of view of personal moral convictions.

The third challenge, the lack of proven methods to translate principles into practice, is interesting from the point of view of SE and is something that has been discussed in existing papers. Various studies (e.g., [4] [6] [5] [12]) argue that principles need to be made more actionable by converting them into SE practices or methods or otherwise repackaging them into some other form. Doing so remains a key challenge in AI ethics, though some steps have nonetheless been taken toward operationalizing these principles.

Morley et al. [6] list various precise technical tools related to the implementation of AI ethics. Ryan & Stahl [12] call for studies to further associate and classify such tools according to the relevant AI ethics principles so as to aid their implementation. While such tools are required to bring AI ethics into ML, AI ethics also needs to be a part of SE [5], given that ML is only a part of the entire software development process [13]. To provide some basis for doing so, we are aware of two methods that have been proposed for implementing AI ethics [9] [8], although they are still arguably far from being 'proven' ways of doing so despite having some empirical support. While various methods for implementing ethics in general exist (Value Sensitive Design methodology etc.), these are not aimed at the specific context of AI ethics and may consequently fail to account for AI/ML-related ethical issues [21]. Other ways of implementing AI ethics could involve ethical user stories [4], for example, but such ideas warrant further study as well.

To summarize, all of these four challenges are ultimately related to the lack of adoption of these guidelines, discussing reasons that contribute towards these guidelines and principles not seeing use in practice. Empirical studies also point towards these guidelines and principles having had little impact on practice thus far [4] [10] [11]. This is an issue given the importance of principles in recent AI ethics discussions. Principles have served as a way of categorizing various AI ethics-related issues (e.g., bias is a common practical issue that falls under the umbrella of the fairness principle). Principles, as established, have also become the primary way of attempting to bring AI ethics into practice.

In the next section, we assume a hypothetical scenario where these principles *are* implemented into a system by the book and review these principles in this context. We aim to understand to what extent doing so would even be possible with the intent to use these principles as presented in AI ethics guidelines.

4. AI ethics as a part of the product

In this section, we explore the implementation of AI ethics principles by reviewing them one principle at a time. We assume the point of view of a finished product. I.e., what would a product that adheres to each of these principles actually look like, and whether it would be possible to accomplish at all? This section involves the eleven most common AI ethics principles listed by Jobin et al. [1], with each subsection discussing one of these eleven principles.

In the interest of space, we have not included the full description for each principle provided by Jobin et al. [1] and instead provide a summary of our own, based on the original description of Jobin et al. [1], before looking at them from the product point of view. Similarly, we briefly list the various aspects of each principle discussed by [1], but only tackle the most relevant aspects of each principle in this paper. While there is something to be said about each aspect, some are arguably more relevant for some systems and less relevant for others – which is also a point we raise in more detail in the discussion. To further emphasize: *the principles discussed here are discussed solely based on Jobin et al. [1]*, as the discussion on the definitions of these principles is still ongoing, and there are conflicts between papers.

For the purposes of this review, we have selected a specific, hypothetical AI system in order to provide us with a clear context in which to evaluate these principles. The system in question is a smart office solution. The goal of the system is to improve employee well-being and productivity, while also providing savings on building upkeep. ML is used in the following ways:

- **Movement tracking.** Cameras and sensors inside the office building are used to understand where employees (do not) spend time inside the building, as well as to understand where positive or negative emotions are often experienced while at work. Used to optimize building layout etc.
- **Facial expression recognition.** Applied to analyze the emotional state of individual employees and to avoid duplicate results in movement tracking.
- **Biometric data.** Gathered through wearable devices (smart rings) in order to evaluate the stress levels and emotional states of the employees. Worn outside work and while working remotely as well.
- **Building upkeep.** Optimizing heating, cooling, lights, etc., to provide savings based on employee activity in the building, among other factors.

The system is used by the management of the companies buying the system, while the employees of these companies (also including the managers themselves) provide the data for the system. The full data set can be accessed by the company whose system it is and who provides the client companies with the system. Company-specific data from one's own company and some anonymous benchmark data can be utilized by the client companies.

4.1. Principles to be implemented

Transparency is the most common AI ethics principle. It includes the concepts of explainability, interpretability, and other acts of communication and disclosure. Discussion on transparency focuses on data use, human-AI interaction, automated decisions, and the purpose of data use or application of AI systems. Increasing the disclosure of information about the following can

increase transparency: use of AI, source code, data use, evidence base for AI use, limitations, laws, responsibility for AI, investments in AI, and possible impact. [1]

Justice and fairness. Justice is mainly expressed in terms of fairness, i.e., through the prevention, monitoring, or mitigation of unwanted bias and discrimination. Rather than being justice in the purely legal sense, it thus also includes social aspects of justice. In more detail, this refers to respect for diversity, inclusion, and equality, the possibility to challenge decisions made by AI, the right to compensation, fair access to AI and data and the benefits of AI, and the impact of AI on the labor market, as well as the need to address democratic or societal issues. Practical key issues are the risk of bias and diversity in data sets. Ways addressing justice and fairness in practice include (1) "technical solutions such as standards explicit normative encoding", (2) transparency, particularly through public awareness, (3) testing, monitoring, auditing, (4) "developing or strengthening the rule of law and the right to appeal, recourse, redress, or remedy", and (5) more diverse development teams and the better inclusion of civil society in an interactive manner. [1]

Non-maleficence refers to avoiding risks or potential harms and is related to safety and security. The system should never cause foreseeable or (un)intentional harm. Harm includes discrimination and violation of privacy in addition to physical, emotional, and economic harm, as well as harm to infrastructure, and harm related to long-term social well-being. More technical ways of implementing non-maleficence include data quality evaluations, emphasise on security and privacy, testing, monitoring, and awareness of the 'dual-use' potential of the system. Governance-level solutions for implementing non-maleficence include active cooperation across disciplines and stakeholders, compliance with existing or new legislation, and the need to establish oversight processes and practices. [1]

Responsibility and accountability. These two principles deal with responsibility in the moral and legal sense. Personal moral convictions are discussed in addition to legal accountability. In discussing accountability, very different actors are discussed: developers, designers, institutions or organizations, and industry. Practical ways of approaching responsibility and accountability include legal liability, the possibility of remedy, identifying reasons and processes that may lead to potential harm, as well as whistle-blowing in case of potential harm, and aiming at promoting diversity. [1]

Privacy in AI ethics is both a value and a right. Privacy issues are most commonly discussed in relation to data (data protection and security), as ML systems rely on large sets of data. Implementing privacy is associated with "differential privacy, privacy by design, data minimization and access control, calls for more research and awareness and regulatory approaches, with sources referring to legal compliance more broadly, or suggesting certificates or the creation or adaptation of laws and regulations to accommodate the specificities of AI." [1]

Beneficence is often discussed as a principle focused on promoting 'good' through human well-being, peace, and happiness, but is seldom accurately defined past this general goal. This includes the creation of socioeconomic opportunities, as well as economic prosperity. Ways of doing so include aligning AI with human values, advancing scientific understanding of the world, minimizing power concentration or using power to advance human rights, working more closely with all possible stakeholders, minimizing conflicts of interests, providing channels and possibilities feedback, and acting on it to prove beneficence, and by creating ways of quantifying human well-being. [1]

Freedom and autonomy focus on both negative and positive freedom in relation to AI. Positive freedom refers to the freedom to flourish, to self-determination through democratic means, the right to establish and develop relationships with other human beings, the freedom to withdraw consent, or the freedom to use a preferred platform or technology. Negative freedom is about, e.g., freedom from technological experimentation, manipulation, or surveillance. Ways to implement freedom and autonomy include focusing on transparent and predictable AI, by not "reducing options for and knowledge of citizens", actively seeking to increase awareness about AI, as well as actively seeking consent from those impacted by it (e.g., data collection). [1]

Trust as a principle is focused on building AI systems the (end-)users and other stakeholders can trust. Trust is often associated with other principles (e.g., transparency, accountability, and reliability), which are considered ways of generating trust. Stakeholders should be able to trust the recommendations or decisions of the AI, as this is imperative for the adoption of the system. Trust is linked to the idea of trustworthy AI. Ways of producing trust include: (1) producing understandable systems, (2) fulfilling public expectations, (3) proof of fairness, and (4) dialogue/stakeholder participation. [1].

Sustainability is about environmental and societal factors. In terms of the environment, it focuses on, e.g., improving ecosystems and biodiversity. As for societal factors, sustainability is about job market factors, fair societies, and the promotion of peace, among other aspects. Ways of achieving sustainable AI include: (1) increasing energy efficiency, (2) minimizing ecological footprint, and (3) ensuring accountability in case of potential job losses. [1]

Dignity is described as a vague and undefined principle. The main point of dignity is to respect the human in the loop. In more detail, this means respecting human rights and otherwise avoiding harm, not forcing acceptance, not automatically classifying individuals, and having no hidden or deceptive human-AI interaction. [1] Practical recommendations for doing so are scarce.

Solidarity stems from the argument that AI systems benefit already well-off individuals and increase inequality. The idea of solidarity includes redistributing the benefits of AI, not threatening social cohesion, and respecting vulnerable persons and groups. [1] It is more focused on the long-term societal effects of AI than on the effects of any single company or its system at present.

4.2. Process and result of applying principles

First, we outlined a hypothetical system to apply the principles discussed above. We outlined goals for the system and what it would look like in practice. After this, we began to go over the principles one at a time, discussing the contents of the principles and the proposed ways of addressing them in practice in the given context. We considered all of these propositions in the system and, in general, outlined a system that would adhere to all "requirements" of each principle outlined in the various guidelines reviewed by [1]. In this fashion, we specified all the aspects of the product that would change as a result of it being developed using AI ethics principles strictly by the book. There were, additionally, various things that should be considered when looking at the larger context that involves the development *process* and the business model of the system, but these were out of the scope of this exercise.

The resulting system is illustrated in Figure 2 (found at the very end of the paper). Figure 2

depicts the ways in which the system was affected by the 11 principles. It also depicts which principles contributed to which feature or part of the system (or related service). In Section 5, we use observations from this thought exercise to illustrate challenges in applying these principles in practice from a product point of view.

5. Discussion

In this section, we discuss our observations from attempting to build a hypothetical system around 11 AI ethics principles. This process and its results are discussed in Section 4. For clarity, we have bolded each main point we are making at the start of the paragraph(s) discussing them.

Potential for information overload. One problem we see with implementing AI ethics directly through principles is a large number of different principles and related concepts associated with each principle. Even for individuals well-versed in the philosophical discussion on the topic, these concepts require studying to internalize, especially when their interplay and the entire body of knowledge on AI ethics is considered. This is arguably even more daunting for developers who may have no prior knowledge of AI ethics or ethics-related training. This is an issue of communication between fields of research and professions and could be, in part, why there has been little interest in AI ethics in SE. Based on this paper, we further highlight the importance of the idea of repackaging and prioritizing principles to make them more actionable or understandable for developers [4] [6] [5]. In this regard, we also feel that implementing ethics directly through guidelines may easily make it a process that is detached from SE activities.

Principles overlap. Another point we wish to raise in relation to a large number of principles and the large number of concepts associated with these principles is that there is indeed some overlap in the principles. This has been pointed out by Morley et al. [6] who, as a solution, recommend focusing on the five principles they highlight in their paper (also found in this paper in Table 1). To some extent, we agree. As we have also illustrated in Figure 2, there is also an overlap between AI ethics principles from a product point of view. To use *privacy* as an example: privacy violations are a part of *non-maleficence*, and being able to exert control over one's own data is a part of *autonomy*. Even on the level of conceptual discussion, having fewer principles that are more accurately defined could be beneficial.

However, for practical implementation, the concept of privacy may feel more tangible. This, again, leads back to the idea of repackaging these principles in some form to make them more approachable in practical use. The authors of methods and other tools for implementing AI ethics need to familiarize themselves with the discussion on AI ethics and make conscious decisions on which principles to include and in what form. This type of debate is at the core of ethics [5].

Principles may conflict in implementation. There are also conflicts between different principles in the same manner. For example, the conflict between beneficence and non-maleficence. Ultimately, it could even be said that the best way to entirely minimize harm (non-maleficence) would be to never develop a system. Yet the principle of beneficence forces one to develop to increase human well-being.

For a more practical discussion, in the case of the example system, the system would be most beneficial if it could be used to handle data about individual employees. This would make it

possible to see if, e.g., individual employees are particularly stressed or tired. However, such information could easily be misused to fire poorly performing employees instead of helping them. Thus, to minimize harm and to prevent the possibility of such unintended use (non-maleficence), the system should not provide (end-)user profiles containing extensive individual data, even in an anonymized form, consequently making it far less useful. While in practice, such situations result in trade-offs, strictly adhering to AI ethics principles leaves less room for them. Similar conflicts are also highlighted by Jobin et al. [1]: "For example, the need for ever larger, more diverse datasets to "unbias" AI appears difficult to conciliate with the requirement to give individuals increased control over their data and its use in order to respect their privacy and autonomy."

Context-specificity of principles. In terms of the practical implementation of principles, the heterogeneity of AI systems also warrants consideration. It arguably depends on the system and its use context, which principles are the most relevant, and how relevant they are. E.g., in our example system *solidarity* had no impact on the system. Principles should be looked at in a context-specific manner to make them more relevant in practice. Aspects to consider include, for example, (1) whether the system is safety-critical, (2) whether the system makes autonomous decisions or simply recommendations, (3) whether the use of the system (or being an object of its data collection) is voluntary in practice, and (4) whether the system handles personal data. While generalizations always have to be made, and it is impossible to account for the uniqueness of each system, creating methods or tooling aimed at certain types or categories of systems according to aspects such as the four we mentioned could be one solution to making these principles more actionable. Reading about the importance of promoting peace or accounting for potential military dual-use may be demotivating and uninteresting for a developer working on an in-house recommendation system for his company. This leads to the following point.

The relevance of high-level global and societal issues. Issues such as promoting peace, accounting for military dual-use, and distribution of wealth are important topics of discussion. However, the best place to do so may not be in tools (e.g., guidelines) meant for *implementing* AI ethics in some form. Doing so may shift focus away from practice and the everyday issues of SE. The role of this discussion is ultimately to raise awareness, but it can be very difficult to convert such issues into requirements or features for most systems.

The relationship between ethics and quality. When approaching ethics from the point of view of the product, it becomes more apparent that some ethical issues are close to issues also associated with quality. Especially the idea of predictability, i.e., that the system works in a predictable manner, as intended and consistently, is also a traditional issue of quality. This relationship between ethics and quality has been explored in the past in SE (e.g., in Agile [22]). However, we argue that quality continues to be a relevant issue in the context of AI. Creating an error-free system is one aspect of creating a system that does not unintentionally cause harm.

In this vein, conventional SE approaches can thus contribute towards realizing AI ethics to some extent. For example, various AI ethics principles stress the importance of stakeholder communication in different ways. While the idea of a stakeholder is understood in a wider sense in AI ethics (e.g., the general public is often considered an important stakeholder) than in SE traditionally, the importance of stakeholder communication and involving (end-)users in

development is a core principle in Agile SE¹. Proper code documentation can similarly contribute towards transparency [19]. This may be something that is occasionally forgotten in the context of ML, however.

Communication is the solution, according to principles and guidelines. As established when discussing overlap between principles, many separate principles advocated for heightened communication with different stakeholders and involving them in the development process (during the operational life of the system as well). This includes both one-way communication from the developers to the users, as well as two-way communication by providing different stakeholders with the means of providing feedback (and receiving compensation).

The importance of transparency is widely acknowledged in AI ethics [1] [6] [18]. As it is not possible to evaluate a system with no knowledge about it, it is considered to enable ethics in the first place. In this regard, there are two issues to tackle: (1) what to communicate and to whom, and (2) how to make sense of the systems themselves to make this communication possible. Disclosure to end-users should be different from disclosure to authorities, for example, as end-users should not be faced with a flood of information they might be inclined to ignore out of convenience.

The second question is how to make sense of the systems themselves. Making understandable AI systems is a technical challenge and not just an ethical one, and requires notable technical expertise. The idea of interpretable or explainable AI systems has been extensively discussed in AI ethics and in ML development as well. This is also a question from the end-user's point of view. While AI ethics principles argue that end-users should be able to understand AI systems (through, e.g., education) [1], is this feasible?

The importance of a service-oriented approach to products in AI/ML systems. While software being a service is now the norm in general, this is further highlighted in AI ethics. As AI ethics principles stress the importance of communication and transparency, many of the practical solutions to these issues include interaction with stakeholders and providing possibilities for doing so. Consequently, many resources need to be devoted towards (end-)user interaction in terms of involvement, feedback channels, providing opportunities for redress, customer support, etc. This unavoidably shifts focus from *product* to *service*.

6. Conclusions

In this paper, we have discussed the current state of AI ethics research with a focus on guidelines and principles. AI ethics research continues to be largely absent in SE, and empirical studies on AI ethics are still scarce. Studies on the state of practice point towards AI ethics receiving little attention out on the field [4], which has been argued to be, in part, due to how difficult it is to implement AI ethics in practice through guidelines and principles.

To encourage more discussion on practical SE in AI ethics, we looked at AI ethics from the point of view of products/services. Utilizing the most prominent AI ethics principles discussed by Jobin et al. [1], we reviewed them in the context of a hypothetical system. We determined ways in which the different principles would affect such a system (or its related services, such as customer feedback possibilities) in practice as features. Based on this review and the current

¹<http://agilemanifesto.org/>

state of AI ethics research, we have discussed potential issues in AI ethics principles and guidelines and suggested steps going forward for AI ethics research.

We summarize our discussion with the following key takeaways:

- Guidelines and principles place notable emphasis on large-scale global and societal issues. AI ethics should focus more on practical development.
- SE methods and other ways of repackaging guidelines into a more actionable form are still needed.
- From the point of view of the product/service, there seems to be a notable overlap in commonly discussed AI ethics principles. When seeking to help implement principles, attention should be paid to potential overlap and conflicts. Conceptual discussion should also further consider such overlap.
- AI/ML systems further emphasize the idea of software products as services. Ethical AI necessitates close cooperation and communication with stakeholders from early design to operations.

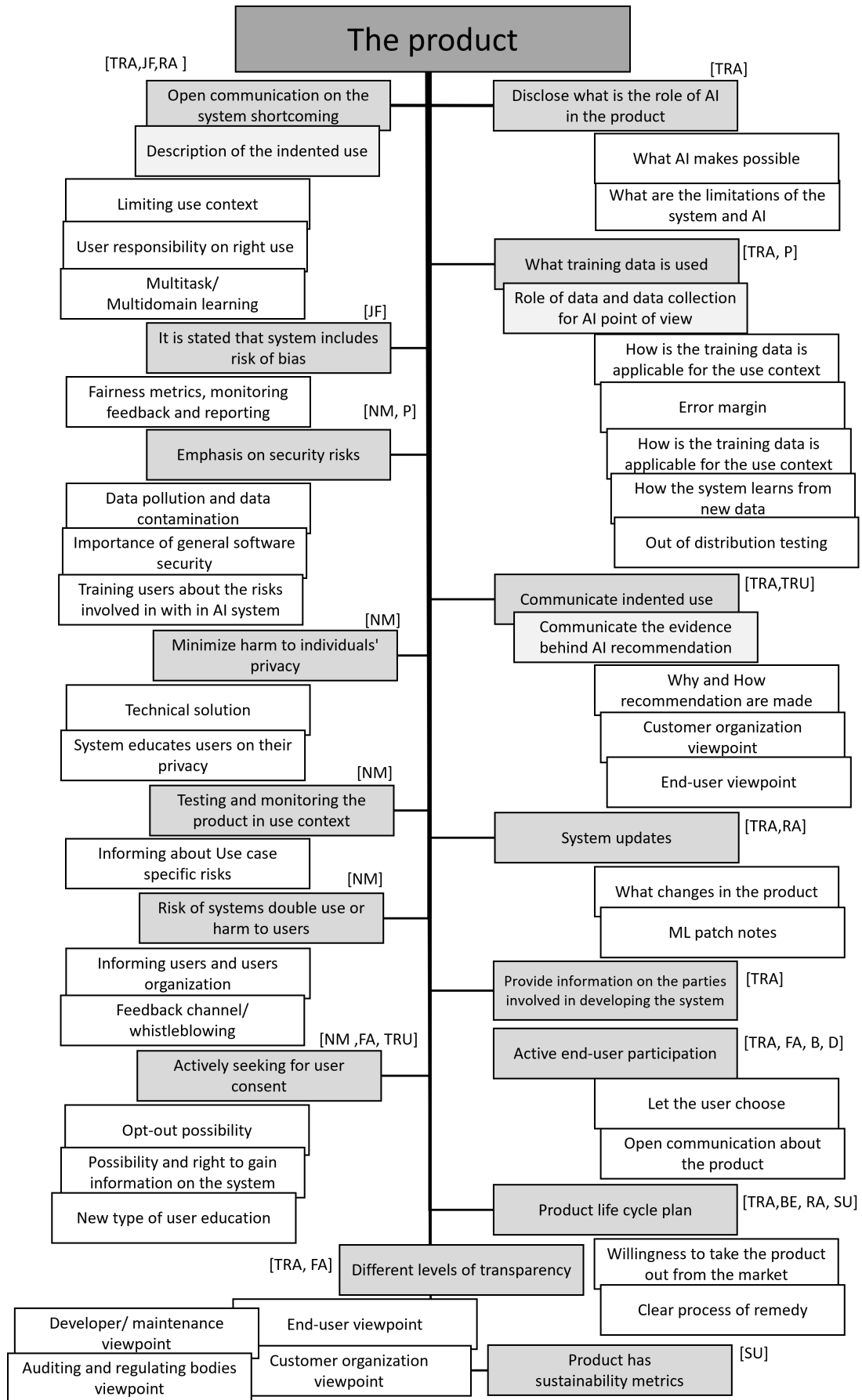
Acknowledgments

This work was partly funded by local authorities (“Business Finland”) under grant agreement ITEA-2020-20219-IML4E of ITEA4 programme.

References

- [1] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, *Nature Machine Intelligence* 1 (2019) 389–399.
- [2] A. McNamara, J. Smith, E. Murphy-Hill, Does acm’s code of ethics change ethical decision making in software development?, in: *Proceedings of the 2018 26th ACM ESEC/FSE, ESEC/FSE 2018*, ACM, New York, NY, USA, 2018, pp. 729–733.
- [3] C. Canca, Operationalizing ai ethics principles, *Communications of the ACM* 63 (2020) 18–21.
- [4] V. Vakkuri, K. Kemell, J. Kultanen, P. Abrahamsson, The current state of industrial practice in artificial intelligence ethics, *IEEE Software* 37 (2020) 50–57.
- [5] B. Mittelstadt, Principles alone cannot guarantee ethical ai, *Nature Machine Intelligence* (2019) 1–7.
- [6] J. Morley, L. Floridi, L. Kinsey, A. Elhalal, From what to how: An initial review of publicly available ai ethics tools, methods and research to translate principles into practices, *Science and Engineering Ethics* 26 (2020) 2141–2168.
- [7] M. Sloane, J. Zakrzewski, German ai start-ups and “ai ethics”: Using a social practice lens for assessing and implementing socio-technical innovation, in: *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’22*, Association for Computing Machinery, New York, NY, USA, 2022, p. 935–947.
- [8] A. P. de Azevedo, H. A. Tives, E. D. Canedo, Guide for artificial intelligence ethical requirements elicitation–re4ai ethical guide, *Proceedings of the 55th Hawaii International Conference on System Sciences* (2022).

- [9] V. Vakkuri, K.-K. Kemell, M. Jantunen, E. Halme, P. Abrahamsson, Eccola—a method for implementing ethically aligned ai systems, *Journal of Systems and Software* 182 (2021).
- [10] V. Vakkuri, K.-K. Kemell, M. Jantunen, P. Abrahamsson, “this is just a prototype”: How ethics are ignored in software startup-like environments, in: V. Stray, R. Hoda, M. Paasi-vaara, P. Kruchten (Eds.), *Agile Processes in Software Engineering and Extreme Programming*, Springer International Publishing, Cham, 2020, pp. 195–210.
- [11] V. Vakkuri, K.-K. Kemell, J. Tolvanen, M. Jantunen, E. Halme, P. Abrahamsson, How do software companies deal with artificial intelligence ethics? a gap analysis, in: *The International Conference on Evaluation and Assessment in Software Engineering 2022, EASE 2022*, Association for Computing Machinery, New York, NY, USA, 2022, p. 100–109.
- [12] M. Ryan, B. C. Stahl, Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications, *Journal of Information, Communication and Ethics in Society* (2020).
- [13] T. Mikkonen, J. K. Nurminen, M. Raatikainen, I. Fronza, N. Mäkitalo, T. Männistö, Is machine learning software just software: A maintainability view, in: D. Winkler, S. Biffi, D. Mendez, M. Wimmer, J. Bergsmann (Eds.), *Software Quality: Future Perspectives on Software Engineering Quality*, Springer International Publishing, Cham, 2021, pp. 94–105.
- [14] T. Granlund, A. Kopponen, V. Stirbu, L. Myllyaho, T. Mikkonen, Mlops challenges in multi-organization setup: Experiences from two real-world cases, in: *2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN)*, 2021, pp. 82–88.
- [15] M. M. John, H. H. Olsson, J. Bosch, Towards mlops: A framework and maturity model, in: *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 2021, pp. 1–8.
- [16] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, D. Dennison, Hidden technical debt in machine learning systems, *Advances in neural information processing systems* 28 (2015).
- [17] L. E. Lwakatare, A. Raj, I. Crnkovic, J. Bosch, H. H. Olsson, Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions, *Information and Software Technology* 127 (2020) 106368.
- [18] T. Hagendorff, The ethics of ai ethics: An evaluation of guidelines, *Minds and Machines* 30 (2020) 99–120.
- [19] V. Vakkuri, K. Kemell, J. Kultanen, M. T. Siponen, P. Abrahamsson, Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study, *arXiv preprint arXiv:1906.07946* (2019).
- [20] D. W. Gotterbarn, B. Brinkman, C. Flick, M. S. Kirkpatrick, K. Miller, K. Vazansky, M. J. Wolf, *Acm code of ethics and professional conduct*, 2018. URL: <https://www.acm.org/code-of-ethics>.
- [21] V. Vakkuri, K.-K. Kemell, Implementing ai ethics in practice: An empirical evaluation of the resolved strategy, in: S. Hyrynsalmi, M. Suoranta, A. Nguyen-Duc, P. Tyrväinen, P. Abrahamsson (Eds.), *Software Business*, Springer International Publishing, Cham, 2019, pp. 260–275.
- [22] H. Abdulhalim, Y. Lurie, S. Mark, Ethics as a quality driver in agile software projects, *Journal of Service Science and Management* 11 (2018) 13–25.



The 11 principles are denoted in the following manner: Transparency = [TRA], Justice and fairness = [JF], Non-maleficence = [NM], Responsibility and accountability = [RA], Privacy = [P], Beneficence = [B], Freedom and autonomy = [FA], Trust = [T], Sustainability = [SU], Dignity = [D], Solidarity = [SO]

Figure 2: Aspects and features to be included to the product