# BUILDING A MATURITY MODEL FOR DEVELOPING ETHICALLY ALIGNED AI SYSTEMS

*Research paper*

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

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

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

Kemell, Kai-Kristian, University of Jyväskylä, Jyväskylä, Finland, kai-kristian.o.kemell@jyu.fi

Rousi, Rebekah, University of Vaasa, Vaasa, Finland; University of Jyväskylä, Jyväskylä, Finland, rebekah.rousi@uwasa.fi

Mikkonen, Tommi, University of Jyväskylä, Jyväskylä , Finland, tommi.j.mikkonen@jyu.fi

Nguyen Duc, Anh, University of South-Eastern Norway, Campus Bø, Norway, angu@usn.no

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

## Abstract

*Ethical concerns related to Artificial Intelligence (AI) equipped systems are prompting demands for ethical AI from all directions. As a response, in recent years public bodies, governments, and companies have rushed to provide guidelines and principles for how AI-based systems are designed and used ethically. We have learned, however, that high-level principles and ethical guidelines cannot be easily converted into actionable advice for industrial organizations that develop AI-based information systems. Maturity models are commonly used in software and systems development companies as a roadmap for improving the performance. We argue that they could also be applied in the context of developing ethically aligned AI systems. In this paper, we propose a maturity model for AI ethics and explain how it can be devised by using a Design Science Research approach.*

*Keywords: Artificial Intelligence, AI, Ethical Alignment, AI Ethics, Maturity Model.*

## 1 Introduction

Artificial Intelligence (AI) technologies are increasingly present as a part of various decision-making processes in healthcare, transportation, urban life, finance and ecommerce (e.g. Panesar, 2019; Sadek, 2007). Consequently, concerns have been raised regarding the ethical impacts of AI Systems (AIS) and their administration (Asaro, 2019; Nowak, Lukowicz, and Horodecki, 2018). While AI offers great benefits, serious hazard may occur as the result of deployment without appropriate assessment of impacts and issues such as accountability and oversight (Whittaker et al., 2018, p. 42).

As stated by Winfield et al. (2019), "robot and AI ethics has been transformed from a niche area of concern of a few engineers, philosophers and law academics, to an international debate" (p. 509). In recent years, AI ethics has been an increasingly popular topic and many studies have been conducted on the ethical consequences of AI systems (Jobin, Ienca, and Vayena, 2019; Ryan and Stahl, 2020). While

most of this discussion has taken place outside Information Systems (IS) venues thus far, we do not consider the topic to be out of scope of IS. This is because AI ethics is closely related to both human-computer interaction (HCI, e.g., how humans interact with AIS) and organizational culture and processes (how organizations develop AIS, preferably ethical AIS). In fact, existing studies on closely related topics also exist in IS literature. In IS, *responsible AI* seems to be one angle taken to AI ethics, as seen in, for example, a recent SJIS call for papers (Scandinavian Journal of Information Systems, 2021). Extant studies on AI ethics in IS have focused on the interaction between humans and intelligent systems (e.g., Amershi, Weld, et al., 2019; Lyytinen, Nickerson, and King, 2020), and on discussions concerning new research perspectives to this topic (e.g., Bailey and Barley, 2020).

As there appears to be gaps to bridge to improve AI ethics practices, a need for more coordination for ethical AI design has emerged (Morley et al., 2020). For example, there is agreement on the need for AI to be ethical, but no consensus on what constitutes ethical AI and what is needed for its realization (Jobin, Ienca, and Vayena, 2019). While theoretical concepts for guiding AI ethics are widely available, translating them into actionable and enforceable practices poses challenges, and many tools created for the purpose are relatively immature (Mittelstadt, 2019; Morley et al., 2020).

Challenges may also arise from the possibility that AI ethics, like ethics in general, lacks "mechanisms to reinforce its own normative claims" (Hagendorff, 2020, p. 99). One of the problems is that establishing policies and enforcing them within an organization remains challenging and unrewarding. Demands for ethical AI are being declared by many, but the rewards of establishing ethical initiatives and commitments remain unclear (e.g. Hagendorff, 2020). When companies and research institutions make "ethically motivated 'self-commitments'" in the AI industry, efforts to formulate a binding legal framework are discouraged, and any demands of AI ethics laws remain relatively vague and superficial (Hagendorff, 2020, p. 100). As Greene, Hoffmann, and Stark (2019) suggest, many high-profile companies, organizations, and communities have signaled their commitment to ethics, but the resulting articulated value statements prompt more questions than answers. It seems that the immediate negative consequences of not applying ethical principles in AI development are not severe enough – or the rewards not lucrative enough – to motivate companies to follow through with ethical principles.

Despite these challenges, several organizations have reacted to ethical concerns relating to AI, for example, by forming ad-hoc expert committees to draft policy documents (Jobin, Ienca, and Vayena, 2019), producing statements that describe ethical principles, values and other abstract requirements for AI development and deployment (Mittelstadt, 2019). At least 84 public-private initiatives promoting AI ethics principles and values were identified by Mittelstadt (2019). Such initiatives are useful for the field as they can "help focus public debate on a common set of issues and principles, and raise awareness among the public, developers and institutions of the ethical challenges that accompany AI" (Mittelstadt, 2019, p. 501).

So far, principles and values used to form *AI ethics guidelines* have been the primary tools intended to help companies develop ethical AI systems. However, guidelines alone cannot guarantee ethical AI systems – additionally, they seem to suffer from a lack of industry adoption (Mittelstadt, 2019; Vakkuri et al., 2020). To close the gap between research and practice, we have started to look into frameworks, models, and other tools that are actively used in the field. All this is in the effort to translate theory to action. As a result, we propose the creation of a maturity model for AI ethics. Maturity models, which we later discuss in detail, are tools for evaluating and improving the level of maturity of organizational processes, often in relation to systems development issues. Similarly, an AI Ethics Maturity Model could help organizations tackle AI ethics issues during AIS development. Maturity models are widely used in the IT industry, and their popularity in the field has partially motivated our approach towards an AI Ethics Maturity Model, as a means to improve practical application of ethics in AI systems development. The maturity model operates through the analogy of a roadmap that indicates where organizations are positioned in order to make it easier to navigate forward in the endeavor to develop ethically aligned AI.

There are already numerous software maturity models (Mettler, Rohner, and Winter, 2010; Poeppelbuss et al., 2011). One question worth asking is whether they could already solve this issue – do we really need an AI Ethics Maturity Model in and of itself? In comparison to traditional non-AI software code,

AI systems are sensitive to some special quality attributes, such as technical debt and anti-patterns, due to various AI-specific issues (Bogner, Verdecchia, and Gerostathopoulos, 2021; Vakkuri, Jantunen, et al., 2021a). While traditional software is deterministic with a predefined test oracle, AI/Machine Learning (ML) models are often probabilistic (Holzinger, Carrington, and Müller, 2020). ML models learn from data and model quality attributes such as accuracy change throughout the process of experimenting (see e.g., Das et al., 2019; Garcez and Lamb, 2020). Moreover, ethical requirements, or attributes such as fairness, trustworthiness, transparency, and explainability (Jobin, Ienca, and Vayena, 2019), have unique meanings in the context of AI, and they are not sufficiently addressed in existing software development models. Moreover, data is the central component of the engineering process with many new problems, such as dealing with missing values, data granularity, design and management of the database, data lake, and the quality of the training data in comparison to real-world data (see e.g., Sambasivan et al., 2021). These differences complicate attempts to apply traditional software development models to AI.

Several AI-specific models have been published, for example, a Microsoft nine-step pipeline (Amershi, Begel, et al., 2019), a five-step "stairway to heaven" AI model (Lwakatare et al., 2019), and a maturity framework for AI process (Akkiraju et al., 2020). However, the focus of these models is not identical to what we want to introduce – they are not especially focused on the quality or ethical aspects of developing AI systems. These models reflect processes in particular organizational contexts, but there is still room for a general model that could be adopted in SMEs and start-up companies (Nguyen-Duc et al., 2020). A more generally applicable AI Ethics Maturity Model is still needed to benchmark and promote the proper engineering practices and processes to plan, implement, as well as integrate ethical requirements. Moreover, this model should facilitate the standardization and dissemination of best practices to developers, scientists and organizations.

This paper extends an earlier workshop paper, "Time for AI (Ethics) Maturity Model Is Now" (Vakkuri et al., 2021b). Here, we describe the theoretical rationale behind our proposed AI ethics maturity model. We turn to Design Science Research as the overarching methodology for the endeavor, while also discussing a framework specific to maturity model development in order to tackle certain parts of the current model development. This endeavor is formulated as a research question:

- *How should an AI Ethics Maturity Model be developed?*

In the rest of the paper, the background of AI ethics is presented, followed by maturity models and the role of culture in their development in Section 2. In Section 3, we introduce the concepts utilized in the maturity model development. In Section 4, the AI Ethics Maturity Model is introduced as it exists in its current state, and then future research steps are deliberated. Finally, in Section 5, we discuss the implications of the AI Ethics Maturity Model.

# 2 Background

## 2.1 AI Ethics

The field studying ethical considerations in AI systems – AI ethics – has been open for initiatives to advance the field (Greene, Hoffmann, and Stark, 2019), and in recent years, many studies have answered the call. The initiatives that have emerged offer varying goals and definitions for what is expected of ethical AI systems. Awad et al. (2018) proposed that we are entering an era where intelligent systems can be tasked "not only to promote well-being and minimize harm, but also to distribute the well-being they create, and the harm they cannot eliminate" (p. 59). To pursue this approach towards understanding and resolving societal consequences, actions are needed: societal and policy guidelines should be established to ensure that intelligent systems "remain human-centric, serving humanity's values and ethical principles" (Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition 2019, p. 2).

We do not have to imagine the negative consequences when something does go wrong with an AI product, since we have already witnessed incidents in which ML based systems learned to exhibit

unethical behavior such as racism or gender discrimination (Vigdor, 2019; Vincent, 2016). These negative outcomes are significant consequences in themselves, as they can negatively impact individuals and society as a whole, but the companies developing these systems have also been affected with bad publicity and were subjects of public criticism for their actions (e.g., Wolf, Miller and Grodzinsky, 2017).

In recent years, as a response to the concerns and discussions around the ethical and societal impacts of intelligent technology, the concept of *AI ethics guidelines*, or principles, has emerged as a common format for communicating ethical intentions. Guidelines for ethical AI development have been published by a variety of organizations, though there appears to be no acknowledged single standard in the field. Guidelines often appear to be either "keyword" style principles such as *accountability* or *transparency* (Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition 2019) or descriptive sentences that present the organization's approach, such as *"We want to develop safe, robust, and explainable AI products"* (Bolle, 2020). The guidelines may serve different purposes for each organization: a corporation's motivation to publish a set of ethical guidelines can be expected to be different from that of a research institution.

As phrased by Fjeld et al. (2020), "seemingly every organization with a connection to technology policy has authored or endorsed a set of principles for AI" (p. 4). This can be seen in some major publications from influential institutions. For instance, the IEEE and the High-Level Expert Group committee appointed by the European Commission, have introduced practical design approaches and suggested standards and principles for ethical AI development and implementation (Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition, 2019; Ethics Guidelines for Trustworthy AI, 2019). Yet, research institutions are only at the tip of the iceberg; a variety of institutions, such as governments and corporations, have taken the initiative to publish their own AI ethics guidelines (Jobin, Ienca, and Vayena, 2019). Even the Vatican has published their contribution, teaming up with IBM and Microsoft to draft a call for AI ethics (Stotler, 2020).

While not legally binding, the effort invested in such guidelines by multiple stakeholders in the field is noteworthy and influential (Jobin, Ienca, and Vayena, 2019). Guidelines can be seen as a "part of a broader debate over how, where, and why these technologies are integrated into political, economic, and social structures" (Greene, Hoffmann, and Stark, 2019, p. 2122). We can witness how guidelines have contributed positively to the development of AI ethics discussion by observing the number of organizations that published their sets of guidelines. Based on the number of organizations that use the common vocabulary of "keyword" guidelines, discussing transparency, fairness, and other such principles, it seems as though guidelines may have developed into a type of "common language" for AI ethics discourse; a familiar format that is easy to adopt and quick to communicate. Researchers have conducted reviews on AI ethics guidelines, considering their implications (e.g. Ryan and Stahl, 2020) and looking for unanimity among them (e.g. Hagendorff, 2020; Jobin, Ienca, and Vayena, 2019). Considering these reviews, certain prevalent guidelines have emerged. For example, Jobin, Ienca, and Vayena (2019) identified a "global convergence emerging around five ethical principles", namely *transparency*, *justice and fairness*, *non-maleficence*, *responsibility*, and *privacy*.

However, guidelines alone do not cater for the whole spectrum of AI ethics challenges. Although some similarities emerge between sources and studies, there is no guarantee of consensus in their application. Even if every organization were to adhere to the exact same set of guidelines, their practical application is not guaranteed to be synchronized. There may be questions related to, for example, interpretation, emphasis and level of commitment, that organizations need to make for themselves. In particular, when considering organizations employing guidelines in their AI product development, the guidelines often provide us with an answer to the question *what* is done, but not *how* (see, e.g., Morley et al., 2020). Additionally, the reliance on guidelines is that their impact on human decision-making is not guaranteed, and they may then remain ineffective (Hagendorff, 2020).

As reported by Vakkuri, Kemell, Kultanen et al. (2019), there appears to be a gap between research and practice in the field of AI ethics when it comes to the procedures of companies. In this context, ethical principles exist but are not acted on, because academic discussions have not carried over into industry

(Vakkuri, Kemell, Kultanen, et al., 2019). It can be argued that developers consider ethics important in principle but may perceive them as too distant from the issues they face within their work. In a survey of industrial practices that included 211 companies, 106 of which develop AI products, it was found that companies have mixed levels of maturity in implementing AI ethics (Vakkuri, Kemell, et al., 2020). In terms of guidelines, the survey discovered that the various AI ethics guidelines had not, in fact, played a notable role in industrial practices, confirming the suspicions voiced by Mittelstadt (2019), who suggested that industry lacks the means to translate AI ethics principles into practice.

The high variety in both industrial practices and AI ethics guidelines may make it difficult to assess AI systems development, especially on aspects such as trustworthiness or other ethics-related topics. To answer the need for standardized evaluation practices, we propose turning our gaze towards maturity models, and their utility in evaluating software development practices. Maturity models, or maturity practices, for AI with different emphases have already been introduced. These can be seen in models such as the AI-RFX Procurement Framework by The Institute for Ethical AI and Machine Learning (The Institute for Ethical AI and Machine Learning, 2020) and the AI Maturity Framework (Ramakrishnan et al., 2020). Next, we discuss maturity models in general, before describing them further in the specific context of AI ethics.

## 2.2 Maturity Models

Maturity models are intended to help companies appraise and develop process maturity. They serve as points of reference for different stages of maturity in a specific area. In the context of software engineering (SE), they are intended to help organizations move from ad-hoc processes to mature and disciplined software processes (Herbsleb et al., 1997). Since the Software Engineering Institute launched the Capability Maturity Model (CMM) almost twenty years ago (Paulk et al., 1993), hundreds of maturity models have been proposed by researchers and practitioners across multiple domains. This has given rise to frameworks for assessing the current effectiveness of an organization and supporting teams in deciphering what capabilities are needed in order to improve their performance.

Though there are numerous maturity models in software development, the Scaled Agile Framework (SAFe) (Scaled Agile, Inc., 2021) and Capability Maturity Model Integration (CMMI) (Paulk et al., 1993) are typical examples of some of the more high-profile maturity models in this area. SAFe is a mixture of different software development practices and focuses mainly on scaling Agile development in larger organizations. CMMI, on the other hand, focuses on improvements related to software development processes. In general, Software Process Improvement tools are rooted in Shewhart-Deming's plan-do-check-act (PDCA) paradigm, where CMMI, for example, represents a prescriptive framework in which the improvements are based on best practices (Pernstål et al., 2019).

Maturity models have also been studied in academic research. Studies have focused on both their benefits and potential drawbacks. For example, a past version of the CMMI has been criticized for creating processes that are too heavy for organizations to handle (Meyer, 2013; Sony, 2019), and in general for being too resource-intensive to adopt in smaller organizations (O'Connor and Coleman, 2009). SAFe, on the other hand, has been criticized for adding bureaucracy to Agile (Ebert and Paasivaara, 2017), leaning towards the waterfall approach.

Nonetheless, these models are widely used in industry, either independently, or in conjunction with other frameworks, tools, or methods. SAFe, for example, has been adopted by 70% of the Forbes 100 companies (Scaled Agile, Inc., 2021). CMMI has even been adopted in fields other than software development. Academic studies aside, companies seem to have taken a liking to maturity models in the context of software. With this in mind, a maturity model for assessing ethics-related AI development maturity could help us convert research knowledge into concrete action in AI ethics.

## 2.3 The Role of Culture in Maturity Model Development

From the culture perspective, we may consider AI ethics, and indeed, efforts to construct and implement AI Maturity Models for application in SE as cultural issues. Here, we bring culture into the discussion

due to the very nature of the construction, manifestation and experience of ethics as cultural discursive practice (Rehg, 1994). That is, our ideas and beliefs of ethics - what is right and wrong, standard, normal or deviant - are formulated by how phenomena are constructed via culture and language. This ties in closely with what we understand as cultural values (Ricoeur, 1973). Values in themselves have been widely recognized in relation to corporate culture, branding and establishing meaningful business-customer relationships (Joyner and Payne, 2002).

Particularly when engaging business with larger global questions and "wicked problems" such as climate change, poverty, immigration, gender etc., ideas of social responsibility and sustainability have been high on the agenda. What must be remembered is that the implementation of ethical practice does not simply exist on ideological, discursive, or communicational levels. Rather, ethical practice is manifested and projected through action, structures and routines. These actions take place as much on the levels of programming and the organizational structures that facilitate programming, as it does at the organizational interface of sales, marketing and service offerings. Therefore, serious consideration needs to be given towards how language and cultural ideas (priorities and beliefs) that connect people to ethical understandings are translated to practical measures that are incorporated into an AI Ethics Maturity Model. There are several ways of observing the relationship between culture and the development of an AI Ethics Maturity Model. One way is the very practical perspective of understanding how culture shapes practice through organization and prioritization, i.e., rapid processes and short-term versus long-term goals (see e.g., Huhtala et al., 2013). Additionally, intentionality of organizational and management culture and the willingness to comply with good ethical practice is another aspect (Verma and Mohapatra, 2020). Another important factor rests on the level of interpretation – translating ideas of ethical conduct into best practice (Clegg, Kornberger, and Rhodes, 2007).

From the perspective of the current research, it is important to gauge how cultural (and organizational cultural) understandings of ethics and their relationship to phenomena, i.e., the programming of intelligent systems, can be deliberated within the development of an AI Ethics Maturity Model. Moreover, and perhaps on a more practical note concerning the implementation, uptake and potential success of such a model, there is the requirement to understand and align the cultures, values, decisions and actions of developers and organizations as a whole towards more common understandings of international ethical practice (Vakkuri, Kemell, Kultanen, et al., 2019; Weller, 2017). For, as previous research and industrial experience have shown (Chow and Cao, 2008), development methodologies such as Agile and any other paradigms of process and production that are introduced into organizational settings, will not gain traction without the support of the organization, its people and cultural properties that enable their acceptance and adoption within everyday routine (Nelson, Taylor, and Walsh, 2014; Weller, 2017).

An AI Ethics Maturity Model incorporates these considerations within its logic as it also assesses the readiness level of the organization, its people, practices and overall operating model for adopting and adapting to the requirements that such guidelines place when implementing ethically aligned SE. If work practices, systems and environments are not established that are able to support developers in their endeavor to align their work with 'best ethical practice' – e.g., too short and unrealistic timelines with too great a task-list; lack of understanding in management; reward for cost and time efficiency, or reward for quality – developers cannot steer their work towards generating better, more sustainable and responsible solutions (Hald, Gillespie, and Reader, 2020).

Ethically aligned AI development practice is greatly reliant on the values of the software company and how it embodies and instils these values within its staff, work conditions, client work, as well as client and general public relationships (Howard, Korver, and Birchard, 2008). This is furthered by the views and beliefs promoted within the companies in relation to how they regard ethics. Many programmers have come forward to mention that they have been surprised by some of the unethical activities their companies have required them to undertake (see e.g. Bort, 2016). Therefore, not only are our understandings of ethics culturally constructed, and dependent on cultural conditions, but adherence to what we understand of ethics through our practice is highly contingent on the conditions within which we operate (Filabi and Bulgarella, 2018). These conditions either enable or disable particular approaches

and practices (Mey and Lloyd, 2016). Similarly, these conditions shape a cultural understanding of ethics in relation to the realities of software development.

# 3 Towards an AI Ethics Maturity Model

So far, we have discussed the reasons why we are proposing an AI Ethics Maturity Model and what we wish to achieve with it. AI systems are becoming increasingly common, and the need for ethical considerations is widely accepted, yet poorly enforced. The field needs standardization as there is a gap between research and practice. In this section, we discuss the building blocks of our proposed AI Ethics Maturity Model to cover the entire sphere of technical and ethical quality requirements. We believe this type of maturity model would help the field move from ad-hoc implementation of ethics – or, even total negligence – to a more mature process level.

Therefore, we believe that a comprehensive AI Ethics Maturity Model should not be an effort for a single researcher or research group, but a multidisciplinary project that builds on a combination of theoretical models and empirical results. The need for this stems from the wish to improve the model's range of applicability as well as the multidisciplinary nature of AI Ethics - a field consisting of disciplines such as philosophy and software engineering. The propositions we make in this paper are intended as a starting point for a comprehensive maturity model for AI ethics. This is an initial framework that will evolve through iteration and we wish to collaborate and receive input from any interested parties. We have initiated research towards an AI Ethics Maturity Model to address the issues considered in this paper. In this section, we introduce a theoretical base for the maturity model, before reporting the current state of our progress in the next section.

## 3.1 Theoretical Building Blocks

De Bruin et al. (2005) stated that, "whilst maturity models are high in number and broad in application, there is little documentation on how to develop a maturity model that is theoretically sound, rigorously tested and widely accepted" (p.2). This appears to still hold true; meta-level research on maturity model building is surprisingly hard to come by, given the number of existing maturity models and their widespread use across industries.

De Bruin et al. (2005) discuss the importance of a standard development framework, in the context of what type of maturity assessment the developed model is intended to undertake: descriptive, prescriptive or comparative. The maturity model we propose is a prescriptive one. A prescriptive model, according to Heldal et al. (2016), is a model that is used to prescribe a not yet existing subject, as opposed to descriptive models, that depict an already existing subject. At this point of research, we are interested in generating a model that provides a tool for assessing ethical maturity; how it *should be* at a certain level, not only how it *is*. The content of a prescriptive model is derived from the information that is available during the time of the model's creation, and according to the specific intent (Heldal et al., 2016). In the case of an AI Ethics Maturity Model, the information we have so far consists of the research we have on AI ethics, maturity model utility, and the current state of AI development practices in organizations. Building on these aspects, the intent of the model, at this point of research, is reliably assessing and improving the maturity of ethical processes in AI development in organizations.

Formulation of requirements for what AI ethics maturity constitutes is an essential step towards creating the model's content. We may require different types of commonly acknowledged agreements on issues that AI maturity entails. We also need to refine a topic still partly shrouded in vagueness – AI ethics – into applicable requirements. The numerous AI ethics guidelines should help in this respect While the existing AI ethics guidelines have faced problems in practical application (Vakkuri, Kemell, et al., 2020), the principles in them are still relevant. Incorporating the principles into a more practical form – such as a maturity model – is what the AI Ethics pursuit is ultimately about. This has already been pursued: the IEEE (Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition 2019) presents an extensive set of guidelines; the EU report Ethics Guidelines for Trustworthy AI (2019) has attempted to make these principles more

actionable. Finally, the ECCOLA method has been designed for practical implementation of AI ethics principles (Vakkuri, Kemell, and Abrahamsson, 2020).

## 3.2 Design and Development Methodology

As we are developing an artefact (maturity model), we turn to Design Science Research (DSR) as the overarching methodology for the endeavor. Specifically, we refer to the DSR Methodology (DSRM) proposed by Peffers, Tuunanen, et al. (2007) as a process to follow during the development of the proposed AI Ethics Maturity Model. This DSRM process, adapted from Peffers, Tuunanen, et al., 2007, and the current state of this endeavor, can be found in Figure 1.

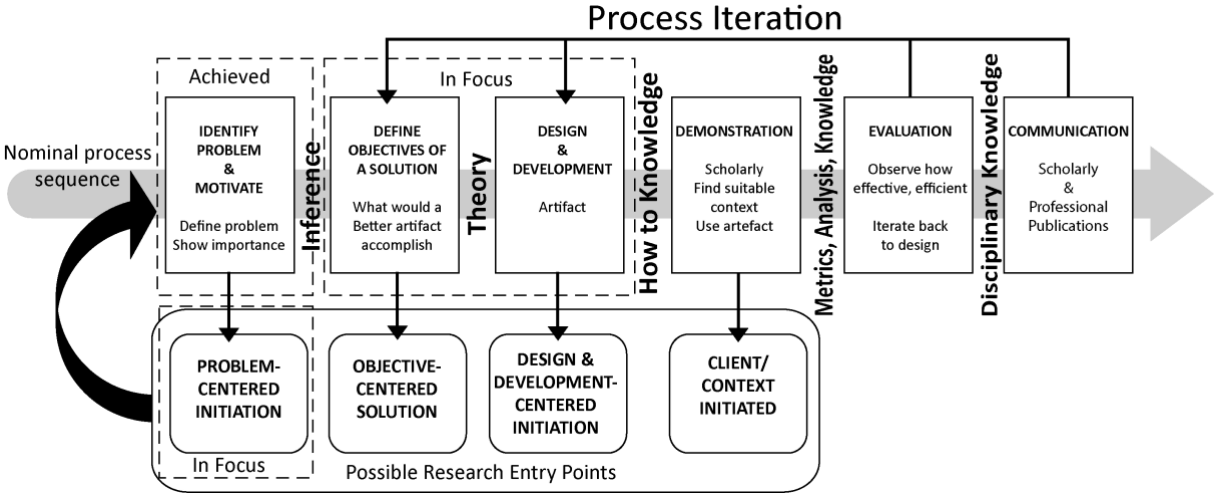


*Figure 1. Positioning the endeavor in the Design Science Research Methodology. Adapted from Peffers, Tuunanen, et al. (2007).*

The DSRM process has been a *Problem-Centric Initiation* where the gap in the area of AI ethics has motivated the development of this current maturity model. As we have discussed previously in this paper, the field heavily relies on guidelines as tools to implement AI ethics in practice (Jobin, Ienca, and Vayena, 2019). Extant research, however, has argued that ethical guidelines do not seem to work in software development in general (McNamara, Smith, and Murphy-Hill, 2018) or in the context of AI ethics (Vakkuri, Kemell, et al., 2020), and thus other approaches, such as maturity models, are needed. The preceding sections of this paper also tackle the *Identify Problem & Motivate* step in general, motivating the model and discussing the problem in detail.

Additionally, we have begun to cover the *Derive Objectives of Solution* step in this paper. We have argued what types of results such a maturity model would ideally achieve over the existing alternatives (primarily guidelines, and some primarily technical, narrowly scoped methods). The exact results that this type of maturity model would achieve in terms of changing organizational processes, however, remain to be determined as the model's development and validation progress.

For the *Design & Development* and *Demonstration* steps of this process, we employ a supporting, context-specific framework especially devised for creating maturity models (De Bruin et al., 2005). In their paper, De Bruin et al. (2005) present a framework for developing maturity models applicable across a range of domains. Their model proposes the following six steps: scope, design, populate, test, deploy, and maintain. We adapt these phases as presented in Figure 2. We discuss this maturity model in the context of this framework in detail in the following section.

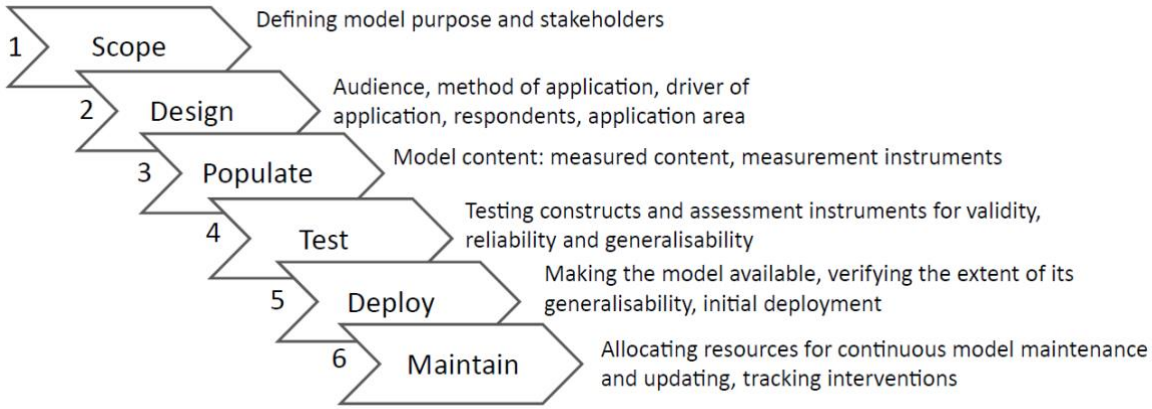


*Figure 2. Development stages adapted from De Bruin et al. (2005).*

Aside from how De Bruin et al. (2005) recommend maturity models be tested, DSR literature should be utilized in the *Demonstration* and especially *Evaluation* phase. Existing IS literature discusses the evaluation of DSR and some frameworks and suggestions for doing so exist. Notable papers on this topic include that of Peffers, Rothenberger, et al. (2012), as well as the paper of Venable, Pries-Heje, and Baskerville (2012).

Given the type of artefact that a maturity model represents, possible evaluation methods, that rely on Peffers, Rothenberger, et al. (2012), would be a Delphi study with AI ethics experts, or a qualitative real-world data approach through Action Research. There may even be other case studies examining the model in action out in the field. Mettler (2011) argues that in DSR, maturity models may undergo either naturalistic or artificial evaluations. For the purposes of this paper we chose a naturalistic evaluation.

Maturity model construction using DSR is a novel area of research. Mettler (2011) discusses maturity model construction through the lens of DSR. In their paper, Mettler (2011) adopts a DSR approach to De Bruin et al.'s (2005) framework in order to better position it within DSR. We take this modified framework into account when discussing the use of De Bruin et al.'s (2005) framework in detail in the next section.

Developing an AI Ethics Maturity Model in this fashion would also contribute to the DSR body of knowledge. Maturity model construction using DSR approaches, to the best of our knowledge, is a novel endeavor. Although Mettler (2011) discuss the construction of a maturity model using DSR, maturity models continue to be largely unexplored artefacts in the context of DSR.

# 4 Development Framework

In this section, we take a more detailed look at the development of the AI Ethics Maturity Model based on the development phases by De Bruin et al. (2005). The methodology underlying the design and development of the maturity model in general was discussed in the preceding section. The initial framework for our AI Ethics Maturity Model is presented in Figure 3. We view the maturity model from the organizational and IS development viewpoints.

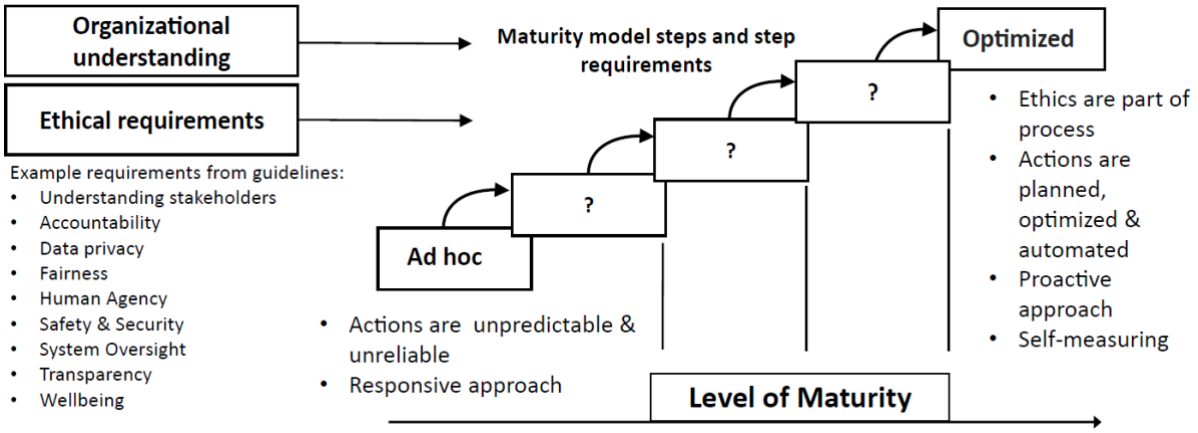


*Figure 3. Initial framework for AI Ethics Maturity Model.*

## 4.1 Development Phases

This section introduces the development stages of the maturity model development according to De Bruin et al. (2005), as introduced in Figure 2. The content of the stages is explained for the context of this maturity model.

**Scope.** One important question spurring the development of our maturity model was whether such a model should be an *AI Ethics Maturity Model* or simply an *AI Maturity Model*. Being a field-specific model, an AI Ethics Maturity Model would address the numerous AI ethics needs discussed in academic literature and public discussion alike. Such a maturity model could be devised so that it would directly complement the on-going principles and guidelines discussion and help bring the model into practice. By focusing on ethics over SE, there is the potential for enhancing suitability for any organization regardless of their chosen development approach. On the other hand, there is always a danger when the model is too focused on AI ethics or design-level issues that focus is drawn away from the technicalities. This could result in a situation where the maturity model would still face issues in practical application, much like the existing guidelines. In general, the model might risk being detached from industry practice. Companies should be closely involved when devising such a model in order to mitigate these potential drawbacks.

The concept of an AI Ethics Maturity Model has grown in relevance, as there is much technical ground to cover when including the whole sphere of AI quality. Implementing ethics into a framework of concretely technical requirements might be problematic, since the combination of technical and ethical requirements might make the model unnecessarily complicated, specific and heavy to apply. Complex conflicts may also emerge: for example, the trade-offs between ethical soundness, effort needed for the technical implementation and financial benefit may end up highlighted as a side product of the model's components - while they could be better addressed as a component of an AI Ethics Maturity Model on their own.

The deliberation process has led to the development of an AI Ethics Maturity Model that incorporates a scope focusing on the assessment of the ethical maturity of processes when developing AI systems. The relevant stakeholders in its development would be a combination of practitioners and academia, since the development process is initiated in academia, but its target users include any organization that develops AI systems. The next steps will include collaboration with industrial partners with a need for AI ethics maturity, to achieve a comprehensive combination of theory and practice. These industrial partners will be engaged through ongoing projects undertaken by the authors.

**Design.** The audience of our AI Ethics Maturity Model is organizations that develop AI systems, and its application purpose is to improve the maturity of ethical practices and practices in AIS development. The stages of maturity (see Figure 3) are not fully defined aside from the beginning and ending step - the model will consist of progressing steps that start from a stage of low maturity and progress to a stage

of high maturity, as made popular by the CMMI and commonly utilized in maturity models. The steps in between will be defined later as the research progresses.

The model's first step, that describes the beginning situation of a company where AI ethics is not considered, is titled *Ad-hoc*. This stage, as the name suggests, implies that ethical demands are reacted to as they arise, and acted upon without a previously conducted plan. Actions are unpredictable, unreliable and unplanned. The last stage, currently called *Optimized* describes a maturity level of proactive approach where ethics are part of the process: actions involving the implementation of ethics are planned, communicated, and effectively measured. At the beginning of the model development, the last stage was titled *Automated*, but during the design iterations it turned out that this name created an unwanted connotation of lacking human agency. Since the model is aiming to find the best solutions from an ethical standpoint, ignoring the human aspect was deemed to be inappropriate.

**Populate.** The intention of the model is to measure the maturity of ethics in AI development processes. Tools and instruments are needed to facilitate the measurement of ethics in AI maturity, that are understood on the basis of the maturity model. Literature reviews have been undertaken to ensure a firm understanding of the state-of-art in the AI ethics field. Through this, we have identified relevant candidate domain components related to *organizational understanding*, such as the unique features of AI technology and the role of culture in the maturity model context - and finally, *AI ethics*, from which we focus specifically on published guidelines. The components contribute to this initial framework of populating the maturity model; the instruments that measure ethical maturity (see Figure 3). The final identification of domain components should result in a larger framework that considers the domain of AI development comprehensively. An extended validation and acquisition process of new domain components could be conducted with, for example, a combination of two methods: surveys to acquire the "big picture" view of the field, and an observational case study of processes in organizations, possibly ones that already consider ethics in a proactive manner.

In our current state of design, we believe that AI ethics guidelines should serve as the core component of the model and be translated into the primary measurement instruments implemented through the model, i.e., the main ethical requirements. For this purpose, the guidelines should be transformed into measurable constructs. In practice, this means devising ways of measuring the extent to which the principles are operationalized. For instance, this may involve assessing the level of transparency in system processes, or how responsibility is distributed and elaborated on in the development. The final framework of the guideline-based instruments is still taking shape as research in the field progresses. There are increasing numbers of guidelines to choose from. We see that culture, as discussed earlier, is the key to identifying the relevant ethical principles for the model as well as dynamic factors affecting both interpretation and implementation of these guidelines. Some principle-related candidates could be, for example, the ethical principles identified by Jobin, Ienca, and Vayena (2019) in their review, or the principles utilized in the ECCOLA method (Vakkuri, Kemell, and Abrahamsson, 2020).

The method of organizing the AI ethics components within the model could include an analysis of the content of each component and rating their urgency in AI systems development. This is in order to construct a hierarchy within the maturity model - from the most essential and urgent components, to the more sophisticated, associative and abstract during latter stages of maturity. The maturity measurements could then be made via methods such as interviews and observations of the organization's actions in relation to development processes, to determine to what extent the conditions of each principle are fulfilled. The components should be validated with expert interviews in the field of maturity models, and possibly a review by academic and professional specialists working in the IS industry - this is one of the key contributions for which we hope to ignite collaboration across the academic sphere of AI research.

**Test.** The testing phase should be carried out with DSR evaluation frameworks in mind (e.g. Venable, Pries-Heje, and Baskerville, 2012). Testing and subsequent evaluation would ideally be carried out in a real-world industrial setting in case companies. Before field tests, expert opinions can be used to further improve the model iteratively.

In the testing phase, when measuring the validity and reliability of our model instruments and constructs, we have to determine which tools or whose expertise will be used to measure the model's success. The model development is at a stage in which planning for testing is not yet possible, as we still have to determine many essential steps that effect it along the way. The tools for testing the validity and reliability of the model instruments and constructs will need decisions on precisely which elements of the development are we measuring. For example, it should be decided if the developed end product should be considered in the maturity measurement, or should the measurements focus strictly on processes during development. We also hope, as presented earlier, that we will receive input from other interested parties before finalizing the model, as we intend for this contribution to eventually lead to a well-rounded AI Ethics Maturity Model that can be applied in a number of domains. Despite the uncertainty around the testing tools, we have a testing environment available in the context of autonomous maritime research, through associated projects.

**Deploy.** In addition to initial tests that are undertaken in artificial settings or that employ evaluation methods such as a Delphi study, data based on real-world evaluations (e.g., case studies) will also be collected. After the initial tests, however, the model would ideally see utilization outside the research setting. This could provide data for validating the model further and improving it based on user experiences from companies.

**Maintain.** At this stage, the model continues to be improved based on real-world data. The model should not be left afloat without guidance or maintenance but should be continuously improved. There should always be someone or something (e.g, a team, an organization) maintaining the model, that is contactable and responsible for the model's utilization and continuous development.

# 5 Discussion and Conclusion

In this paper, we laid the foundation for an AI Ethics Maturity Model design based on the development framework by De Bruin et al. (2005), its adaptation by Mettler (2011), and DSRM as the overarching methodology for the endeavor (Peffers, Tuunanen, et al., 2007). We started out with the research question: how should an AI Ethics Maturity Model be developed? We examined the issues present in the field of AI systems development. In particular, we would like to discuss a vagueness and shortage of enforcement of ethical practices in AI systems development. To remedy that, we introduced our proposition of an AI Ethics Maturity Model. This is a model for improving the maturity and subsequent readiness of ethical practices in AI systems development. We argue that the ethical considerations of AI are no trivial issues, as the impacts of AI systems are significant. Yet, as researchers have discovered, they are not consistently considered in the practice of AI development. Moreover, due to the contextual and cultural relativity of ethical understandings, the lack of consensus on the conceptual level poses great challenges on practical levels, from coding tasks to organizational culture. Consequently, devising a concrete AI ethics maturity level requires a deep systemic overview of the factors enabling and hindering ethical AI practice. This firm connection between cultural levels, AI ethics and indeed maturity model development appears to be somewhat overlooked in the field of design science research.

The design of the model's maturity stages at this point is adapted from the CMMI model, due to its success as a process improvement tool for software companies. The model's components that measure AI ethics maturity are built upon a foundation of our framework of elements drawn from a literature review. The core components have been formed by a set of ethical principles extracted from AI ethics guidelines. While technical models such as the CMMI model are useful from the maturity model development process perspective, we feel that more work needs to be done in terms of systematically incorporating culture and its dimensions as enablers or disablers within the AI Ethics Maturity Model development. To start with, organizational culture is already integral to maturity model development which renders close cooperation with organizations necessary. Yet, given the high level of cultural dependency within issues surrounding ethics, strong structural connections need to be made between cultural interpretations and practice in organizations and development teams, and how these are tangibly applied to code.

We believe that the maturity model development requires a larger effort than one research group alone can commit to. Subsequently, we intend to collaborate with other interested parties in developing an academically sound, industry-validated model for ethical AI maturity with a comprehensive set of technical, cultural and organizational considerations.

As for the implications of the model, we argue that our proposed scientifically multidisciplinary-based maturity model will lead to a positive contribution in the IS field. The model offers tools to improve processes in different layers of development through ethical principles and their multidimensional connections. For example, the operation of a maturity model in itself creates transparency for the organization as all processes, with actions, are described, executed and verified. We believe that applying the model will contribute to development practices that will improve the quality of AI systems as a result. This in turn, will contribute to the development of more sustainable AI systems based on positive social and ethical impact.

## References


Akkiraju, R., V. Sinha, A. Xu, J. Mahmud, P. Gundecha, Z. Liu, X. Liu, and J. Schumacher (2020). "Characterizing Machine Learning Processes: A Maturity Framework." In: International Conference on Business Process Management. Springer, pp. 17–31.

Amershi, S., A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann (2019). "Software engineering for machine learning: A case study." In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). IEEE, pp. 291–300.

Amershi, S., D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz (2019). "Guidelines for Human-AI Interaction." In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, pp. 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300233. URL: https://doi.org/10.1145/3290605.3300233.

Asaro, P. M. (2019). "AI ethics in predictive policing: From models of threat to an ethics of care." IEEE Technology and Society Magazine 38 (2), 40–53.

Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan (2018). "The moral machine experiment." Nature 563 (7729), 59–64.

Bailey, D. E. and S. R. Barley (2020). "Beyond design and use: How scholars should study intelligent technologies." Information and Organization 30 (2), 100286.

Bogner, J., R. Verdecchia, and I. Gerostathopoulos (2021). "Characterizing Technical Debt and Antipatterns in AI-Based Systems: A Systematic Mapping Study." arXiv preprint arXiv:2103.09783.

Bolle, M. (2020). Code of ethics for AI. Tech. rep. Robert Bosch GmbH. URL: https://www.bosch.com/stories/ethical-guidelines-for-artificial-intelligence/.

Bort, J. (2016). Programmers confess unethical, illegal tasks asked of them to do. URL: https://www.businessinsider.com/programmers-confess-unethical-illegal-tasks-asked-ofthem-2016-11?r=US&IR=T.

Chow, T. and D.-B. Cao (2008). "A survey study of critical success factors in agile software projects." Journal of systems and software 81 (6), 961–971.

Clegg, S., M. Kornberger, and C. Rhodes (2007). "Business ethics as practice." British Journal of Management 18 (2), 107–122.

Das, M., D. S. Dhami, G. Kunapuli, K. Kersting, and S. Natarajan (2019). "Fast relational probabilistic inference and learning: Approximate counting via hypergraphs." In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 01, pp. 7816–7824.

De Bruin, T., M. Rosemann, R. Freeze, and U. Kaulkarni (2005). "Understanding the main phases of developing a maturity assessment model." In: Australasian Conference on Information Systems (ACIS): Australasian Chapter of the Association for Information Systems, pp. 8–19.

Ebert, C. and M. Paasivaara (2017). "Scaling agile." Ieee Software 34 (6), 98–103.

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

Ethics Guidelines for Trustworthy AI (2019). URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Filabi, A. and C. Bulgarella (2018). "Organizational culture drives ethical behaviour: Evidence from pilot studies." In: Anti-Corruption & Integrity Forum.

Fjeld, J., N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar (2020). "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI." Berkman Klein Center Research Publication 2020-1.

Garcez, A. d. and L. C. Lamb (2020). "Neurosymbolic AI: The 3rd Wave." arXiv preprint arXiv:2012.05876.

Greene, D., A. L. Hoffmann, and L. Stark (2019). "Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning." In: Proceedings of the 52nd Hawaii International Conference on System Sciences.

Hagendorff, T. (2020). "The ethics of AI ethics: An evaluation of guidelines." Minds and Machines, 1–22.

Hald, E. J., A. Gillespie, and T. W. Reader (2020). "Causal and corrective organisational culture: a systematic review of case studies of institutional failure." Journal of Business Ethics, 1–27.

Heldal, R., P. Pelliccione, U. Eliasson, J. Lantz, J. Derehag, and J. Whittle (2016). "Descriptive vs prescriptive models in industry." In: Proceedings of the acm/ieee 19th international conference on model driven engineering languages and systems, pp. 216–226.

Herbsleb, J., D. Zubrow, D. Goldenson, W. Hayes, and M. Paulk (1997). "Software quality and the capability maturity model." Communications of the ACM 40 (6), 30–40.

Holzinger, A., A. Carrington, and H. Müller (2020). "Measuring the quality of explanations: the system causability scale (SCS)." KI-Künstliche Intelligenz, 1–6.

Howard, R. A., C. D. Korver, and B. Birchard (2008). Ethics for the real world: Creating a personal code to guide decisions in work and life. Harvard Business Press.

Huhtala, M., T. Feldt, K. Hyvönen, and S. Mauno (2013). "Ethical organisational culture as a context for managers' personal work goals." Journal of Business Ethics 114 (2), 265–282.

Jobin, A., M. Ienca, and E. Vayena (2019). "The global landscape of AI ethics guidelines." Nature Machine Intelligence 1 (9), 389–399.

Joyner, B. E. and D. Payne (2002). "Evolution and implementation: A study of values, business ethics and corporate social responsibility." journal of Business Ethics 41 (4), 297–311.

Lwakatare, L. E., A. Raj, J. Bosch, H. H. Olsson, and I. Crnkovic (2019). "A Taxonomy of Software Engineering Challenges for Machine Learning Systems: An Empirical Investigation." In: Agile Processes in Software Engineering and Extreme Programming. Ed. by P. Kruchten, S. Fraser, and F. Coallier. Cham: Springer International Publishing, pp. 227–243.

Lyytinen, K., J. V. Nickerson, and J. L. King (2020). "Metahuman systems= humans+ machines that learn." Journal of Information Technology, 0268396220915917.

McNamara, A., J. Smith, and E. Murphy-Hill (2018). "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. Lake Buena Vista, FL, USA: ACM, pp. 729–733. ISBN: 978-1-4503-5573-5. DOI: 10.1145/3236024.3264833.

Mettler, T. (2011). "Maturity assessment models: a design science research approach." International Journal of Society Systems Science 3 (1-2), 81–98.

Mettler, T., P. Rohner, and R. Winter (2010). "Towards a classification of maturity models in information systems." In: Management of the interconnected world. Springer, pp. 333–340.

Mey, M. R. and H. R. Lloyd (2016). "Ethics and Organizational Culture." In: Global Encyclopedia of Public Administration, Public Policy, and Governance. Ed. by A. Farazmand. Cham: Springer International Publishing, pp. 1–8. ISBN: 978-3-319-31816-5. DOI: 10.1007/978-3-319-31816-5_2459-1. URL: https://doi.org/10.1007/978-3-319-31816-5_2459-1

Meyer, B. (2013). What is wrong with CMMI. URL: https://bertrandmeyer.com/2013/05/12/whatis-wrong-with-cmmi/.

Mittelstadt, B. (2019). "Principles alone cannot guarantee ethical AI." Nature Machine Intelligence, 1–7.

Morley, J., L. Floridi, L. Kinsey, and A. Elhalal (2020). "From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices." Science and engineering ethics 26 (4), 2141–2168.

Nelson, W. A., E. Taylor, and T. Walsh (2014). "Building an ethical organizational culture." The health care manager 33 (2), 158–164.

Nguyen-Duc, A., I. Sundbø, E. Nascimento, T. Conte, I. Ahmed, and P. Abrahamsson (2020). "A Multiple Case Study of Artificial Intelligent System Development in Industry." In: Proceedings of the Evaluation and Assessment in Software Engineering. EASE '20. Trondheim, Norway: Association for Computing Machinery, pp. 1–10. DOI: 10.1145/3383219.3383220. (Visited on 04/27/2020).