Scandinavian Journal of Information Systems

Volume 34 | Issue 2 Article 6

12-31-2022

Exploring tensions in Responsible AI in practice. An interview study on AI practices in and for Swedish public organizations

Clàudia Figueras
Stockholm University, claudia@dsv.su.se

Harko Verhagen
Stockholm University, verhagen@dsv.su.se

Teresa Cerratto Pargman
Stockholm University, tessy@dsv.su.se

© Scandinavian Journal of Information Systems, 2022 34(2), 199-232

Exploring Tensions in Responsible AI in Practice

An interview study on AI practices in and for Swedish public organizations

Clàudia Figueras, Harko Verhagen, Teresa Cerratto-Pargman
Stockholm University, Sweden
claudia@dsv.su.se
verhagen@dsv.su.se
tessy@dsv.su.se

Abstract. The increasing use of Artificial Intelligence (AI) systems has sparked discussions regarding developing ethically responsible technology. Consequently, various organizations have released high-level AI ethics frameworks to assist in AI design. However, we still know too little about how AI ethics principles are perceived and work in practice, especially in public organizations. This study examines how AI practitioners perceive ethical issues in their work concerning AI design and how they interpret and put them into practice. We conducted an empirical study consisting of semi-structured qualitative interviews with AI practitioners working in or for public organizations. Taking the lens provided by the In-Action Ethics framework and previous studies on ethical tensions, we analyzed practitioners' interpretations of AI ethics principles and their application in practice. We found tensions between practitioners' interpretation of ethical principles in their work and ethos tensions. In this vein, we argue that understanding the different tensions that can occur in practice and how they are tackled is key to studying ethics in practice. Understanding how AI practitioners perceive and apply ethical principles is necessary for practical ethics to contribute toward an empirically grounded, Responsible AI.

Keywords: Responsible AI, AI ethics in practice, empirical studies on ethics, ethos tensions, AI practitioners.

Accepting editor: Polyxeni Vassilakopoulou

1 Introduction

Artificial Intelligence (AI) systems, specifically machine learning systems, are increasingly applied in workflows at many public institutions. AI systems automate tasks such as facial recognition, pretrial and sentencing risk assessment, machine translation, or spam filtering. In this way, AI systems gradually become pervasive and invisible to us (Susser, 2019). Although these systems can potentially benefit people, they can also be harmful, especially to individuals and groups with social, political, cultural, gender, and economically disadvantaged identities (Crawford & Calo, 2016). A central concern for public organizations working for civil society is identifying risks and potential harms with AI in development, design, and deployment practices (henceforth shortened as 'AI design'). The implications of working with biased algorithms in the public sector can be detrimental to society, as underscored by (de Vries, 2020).

Definitions of AI systems are manifold. This study uses Bellman's definition of AI, namely technologies that automate tasks associated with human thinking, such as decision-making, problem-solving, or learning (Bellman, 1978; Russell & Norvig, 2016, p. 2). However, current AI systems display intelligence in limited domains, and the definition of what is considered 'intelligent' is also criticized (for a more extended discussion, see (Cave, 2020). In this study, we focus on the broader impacts that (fully or partially) automated decision-making processes have on society, not on specific technical features of AI. More precisely, we consider the notion of *Responsible AI* by Virginia Dignum (2019) that reintroduces the role of social context and human values in discourses about AI design. Responsible AI addresses the AI-driven harms emphasized by AI ethics guidelines and frameworks and aims to assist in designing ethically responsible AI systems (Jobin et al., 2019). Other examples of such frameworks are Trustworthy AI (such as the EU *Ethics guidelines for Trustworthy AI,* which are later covered in this document), Ethical AI (Winfield et al., 2019), or the Ethically Aligned Design (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019). Despite their different backgrounds, these frameworks address similar core ethical issues: transparency, justice, non-maleficence, responsibility, and privacy (Jobin et al., 2019).

While ethical frameworks are necessary to discuss ethical considerations, they do not by themselves change unethical behavior unless embedded in organizational culture and actively enforced (Mittelstadt, 2019). As such, responsible AI design needs to consider the societal context, human principles, and values (Dignum, 2019). To design AI responsibly, Dignum argues that we need to study ethics *in* Design, which refers to the governing and technical processes supporting the design and evaluation of AI systems to guarantee that the principles of accountability, responsibility, and transparency are at the core of the AI design. The ethics *in* Design can make AI practitioners aware "of the

potential consequences for individuals and societies, by anticipating the consequences of the design choices, reflecting upon the problem being solved by engaging all stakeholders, verifying and validating the design, and taking appropriate action to ensure social, legal and ethical acceptability of the system". (Dignum, 2019, p. 6). Responsible AI then means that AI systems must be recognized as part of a complex sociotechnical system that requires an empirical ethics approach (Dignum, 2019). Currently, empirical studies focusing on empirical ethics in AI design are scarce (Morley et al., 2020), especially those focused on how ethical decision-making is enacted[1] by people behind AI design, namely the AI practitioners.

Tensions, misalignments between ethical principles, or between those principles and practitioners' interpretations may arise when designing AI systems. For example, implementing transparency may unintendedly collide with privacy breaches. Nevertheless, tensions can be "an important way of bridging the gap between abstract ethical principles and specific cases, and therefore an important first step towards an ethics of AI that is practical and action-guiding". (Whittlestone et al., 2019, p. 197). Furthermore, other tensions may occur, such as between different views among co-workers and organizations or when enacting such principles in practice. Therefore, this study investigates how AI practitioners perceive and enact ethics in practice to understand how and where tensions occur. We thus designed an exploratory study to examine the following research questions:

How do AI practitioners perceive ethical issues in their work regarding AI design, and how do they interpret and enact ethical principles in practice?

Decisions made by AI practitioners in high-stakes domains, such as in the public sector, have profound social impacts on shaping the core aspects of AI systems (Martin et al., 2020). Research on AI practitioners typically samples from private companies (especially Big Tech) and academia because these groups are more accessible (Hopkins & Booth, 2021). Consequently, other communities, such as the public sector, remain understudied. In the public sector, IT projects are usually resource-constrained and cross scales and chains of accountability, making them complex to manage and maintain (Veale et al., 2018). Moreover, public agencies' legitimacy depends on public trust; thus, there are different incentives for effective and accountable delivery of services to the general public than private actors who rely on profit-making (Ada Lovelace Institute et al., 2021). Recently, government agencies across Europe have progressively deployed AI and automated decision-making systems in their workflows (AlgorithmWatch, 2020). In Sweden, the government invests in placing the country at the fore-

front of AI development (Government Offices of Sweden, 2019). For all these reasons, we choose to interview AI practitioners that work for the public sector in Sweden. This study addresses AI practitioners who work in or for the Swedish public sector, specifically for government agencies[2] that directly interact with the general public, such as the tax authorities.

We designated AI practitioners as the people who are behind the development, design, strategy, research, and management of AI systems. We drew upon semi-structured interviews that referred to the EU *Ethics guidelines for Trustworthy AI* as the official document reflecting a common and normative understanding of AI ethics to examine the ethical understanding by AI practitioners. The document worked as a prompt for the interviewees to engage with ethics in the conversation. We selected the EU Ethics guidelines for Trustworthy AI because we expected our informants, who are located in Sweden, to be more familiar with the EU Ethics guidelines for Trustworthy AI than with another Responsible AI framework.

The selected qualitative methodological approach has supported us in building a rich context around the people behind designing AI. Interviews helped create the space for participants to speak openly about their perceptions and values, allowing us to ask follow-up questions and get unexpected findings. Moreover, engaging in conversations with the practitioners about their everyday practices increases the understanding of aspects that may not be evident from top-down approaches or documents (Veale et al., 2018).

Our findings contribute to unpacking how AI practitioners who work with AI in public organizations in Sweden apply and perceive ethical considerations in their practices. Such knowledge is vital to gaining a grounded and empirically understanding of how Responsible AI practices unfold in projects conducted in the public sector. In particular, we identified tensions in the practitioners' practice by conducting and analyzing the interviews. Analyzing these tensions contributes to understanding how practitioners deal with ethical principles in AI design practice.

The article is structured as follows. First, we provide the background on how previous research addressed the study of ethics of AI in practice. Moreover, we present the EU *Ethics guidelines for Trustworthy* AI, which we use as the common ground to prompt discussions on ethics in the interviews. Following this, we delve into the relevant conceptual frameworks that were the basis of our study, namely In-Action Ethics and tensions in design. After that, the research methods are elaborated. Finally, we present the analysis of the findings and discuss them along with potential future research directions.

2 The gap between AI ethics principles and AI design practices

Discourses about responsibility and AI ethics are predominantly centered on policy documents (such as the EU guidelines) and conceptual work about how ethics *should* be applied in all sectors rather than how ethics *is* applied in specific sectors. Thus, ethics is not always viewed as linked to the particular context of the sociotechnical systems or compliance with the AI ethics guidelines. Although much research on AI Ethics has been conducted in conceptual terms, a growing body of literature reflects a sustained interest in contributing to discussions on Responsible AI (Dignum, 2019) in research communities such as ACM Conference on Fairness, Accountability, and Transparency (FAccT) or the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), and the like.

Among the few studies which engage with the intricacies and messiness of ethical considerations concerning AI systems in organizations, we find Wang et al. (2020). They suggest five organizational strategies for firms to adopt responsible AI practices, including creating socially accountable strategies and mechanisms to regulate AI usage. Schiff et al. (2021) conceptually analyze the literature on the inherent challenges of applying Responsible AI guidelines in private organizations in the USA. The authors describe six contributing factors to the gap between the ethical principles and professional practices: overabundance of tools, accountability distribution problem, sociotechnical disciplinary divides, the complexity of AI's impacts on well-being, and incentives dilemma. These suggestions are supported by interviews with industry AI practitioners by Rakova et al. (2020). These authors found the interviewees struggle with a lack of accountability due to uncertainty about where responsibility falls, a lack of support to conduct responsible AI (including inadequate credit and insufficient compensation for their impact), and misalignment of incentives between individuals, teams, and organizations. They conclude their study by highlighting that to study the effects of AI systems, it is crucial to consider the people building them and the organizations' structure and culture. Similarly, Seppälä et al. (2021) conducted expert interviews in (primarily private) organizations deploying AI systems to empirically elucidate how principles of ethical AI are translated into organizational practices. Mayer et al. (2021) conducted expert interviews in private companies to identify mechanisms corporations use to encourage ethical AI practices. Both studies found that AI governance practices such as risk assessment and cross-functional collaboration are essential to engage in ethical AI. All these studies cited so far focused on AI design in private organizations.

Among the empirical studies on AI design in the public sector, Veale et al. (2018) interview study with public sector AI practitioners calls for studying ethics "*in vivo*, in

the messy, sociotechnical contexts in which they inevitably exist," including the broader institutional and political contexts (Veale et al., 2018, p. 10). Relatedly, Morley et al. (2020) note that most developers long for practical resources such as tools and methods to help them and are frustrated by these abstract principles. The available tools also hardly help the developers consider the users' autonomy and those affected by the developed systems (Morley et al., 2020). Moreover, several studies pointed out that the available AI ethics tools (i.e., ethical guidelines, frameworks, and models) are challenging to use in practice (Vakkuri et al., 2020; Jantunen et al., 2021). Such tools do not usually support integrating ethical thinking throughout artifact design, and usually, ethical considerations are put forward after the initial design phases. Integrating ethical aspects into well-developed software engineering practices and processes, from coding tasks to organizational culture, is a challenge (Vakkuri et al., 2020; Jantunen et al., 2021). Jantunen et al. (2021) argue that this may be due to the cultural and contextual relativity of ethical understandings and the lack of consensus on the conceptual level and suggest that stronger structural connections must be made "between cultural interpretations and practice in organizations and development teams" (Jantunen et al., 2021, p. 12).

2.1 The EU guidelines

The European Commission appointed in June 2018 the independent High-Level Expert Group on Artificial Intelligence (AI HLEG). One of their deliveries is the *Ethics guidelines for Trustworthy AI* (AI-HLEG, 2019) (hereafter shortened as 'the Guidelines'). The Guidelines contain seven key requirements based on four ethical principles (respect for human autonomy, prevention of harm, fairness, and explicability) for Trustworthy AI: (1) *human agency and oversight,* (2) t*echnical robustness and safety*, (3) *privacy and data governance*, (4) *transparency*, (5) *diversity, non-discrimination and fairness*, (6) *environmental and societal well-being,* and (7) *accountability*. For this study, we have paid particular attention to the following principles: *transparency*, *fairness*, and *stakeholder consideration and involvement* (found within requirements (5) and (6)). These principles reflect the sociotechnical nature of AI systems implemented in public organizations. We expected practitioners to have less formalized procedures to deal with these principles compared to other more technical requirements such as (2) or (3).

Transparency is linked with the principle of explicability. This principle is listed in the Guidelines with fairness as one of the four main ethical principles for Trustworthy AI, and it is also a key requirement by itself. Explicability refers to openly reporting the

abilities and motivations of AI systems. Besides, the decisions taken by those systems should be explainable as much as possible to those directly or indirectly affected. Thus, explicability becomes essential to build users' trust in AI systems. In those cases where algorithms consist of black boxes, alternative explicability measures such as auditability, traceability, and transparent communication on the system's capabilities should be conveyed.

The Guidelines list the principle of *fairness* as one of the four ethical principles rooted in fundamental rights that must be respected to guarantee that AI systems are developed, deployed, and used responsibly. Although multiple interpretations of fairness exist, the Guidelines assert that fairness has a substantive and a procedural dimension. The substantive dimension implies avoiding unfair bias, discrimination, and stigmatization, equal opportunity of access, and just distribution of benefits and costs. The procedural dimension includes the ability to contest an outcome and pursue redress against decisions made by the AI system and the humans operating them. Fairness belongs to the key requirement of diversity, non-discrimination and fairness, which includes avoiding unfair bias, aiming toward accessibility and universal design, and stakeholder participation.

To achieve Trustworthy AI, the Guidelines set forth that we must *consider and involve all affected stakeholders* throughout the entire AI system's life cycle. This requisite is linked to the principle of fairness and prevention of harm. It can be found partially in the key requirements of *diversity, non-discrimination and fairness,* and *societal and environmental well-being*. Even after deployment, the Guidelines strongly recommend requesting feedback from stakeholders directly or indirectly affected by the AI system (including the public, private and non-profit sectors) and creating mechanisms that allow participation. This goes beyond eliciting feedback from users or stakeholders to build the proper user experience (common in Scrum and Agile software development), as the recommendation is also to include indirectly affected people and raise their voices to create meaningful participation.

3 Theoretical background on ethics

3.1 Ethics of technology

Ethics can be seen as the discipline concerned with the study of morality and expressed in how we should live, what a good life consists of, how we should treat others, and the kind of society we want to live in (Frauenberger et al., 2017). Ethics in academic

contexts tend to be divided into theoretical ethics or meta-ethics, applied ethics, and normative ethics. However, the theorization of ethics has been going through developments in the last forty years: from applying ethics in specific fields (e.g., medicine, engineering), feminist critique to normative ethical theories (ethics of care) to empirical ethics (focusing on the context and participants) (Frauenberger et al., 2017).

Concerning the ethics of technology, we follow Verbeek (2006), who argues that the products of technology, i.e., artifacts, structure the perception of reality and, thus, what is real. This structuring of reality by technology, in turn, influences the ethical decision-making in this reality, giving designers responsibility for the results of using their artifacts. As he puts it: "Engineering design is an inherently moral activity" (Verbeek, 2006, p. 368). He distinguishes between the context of design and the context of use which is of great importance for involving all stakeholders, direct and indirect. On that note, Devon and van de Poel (2004) propose a social ethics approach to study the crucial importance of the social context of making ethical decisions and suggest the ethical aspects to be necessary during the whole life-cycle of the artifacts. During this process, learning of the values and their role can occur. These works reflect on the artifact's consequences and the social context, the context of design and use, the involvement of all stakeholders, and the balancing of the general and particular approaches. This ongoing process of reflection throughout the product life cycle generates learning about values, refining the practitioners' ethical compass or ethos.

3.2 Computer ethics and AI ethics

The applied ethics concerning IT is usually named c*omputer ethics* and is aimed at ethical issues related to the use and development of computer artifacts. Although the discussion of the ethical aspects of computing is often seen as relatively new, it was already part of the birth of modern computer science around the 1940s. Even some of the issues at stake are reminiscent of the current ethical challenges, such as automatic radar identification, which excluded human involvement from the decision-making (Bynum, 2008).

Computer ethics starts from understanding the values at stake and converting these into concrete consequences for the design process. However, a recent study found that human values were only directly considered in a few published studies in software engineering conferences or journals (Perera et al., 2020). From empirical investigations, it is known that about half of IT engineers were not convinced they were responsible for considering a value such as privacy when developing or implementing systems

(Spiekermann et al., 2019) and that ethical responsibility is seen as a matter of the legal department (Bednar et al., 2019).

According to Prior et al. (2002), the ethical attitudes of information systems professionals found that codes of ethics have a negligible effect compared to this denial of responsibility. Their study suggests a list of potential measures to increase the awareness and handling of ethical issues. These include clear policies regarding some of the ethical challenges addressed in the survey (using computing resources, respecting intellectual property, among others), training newcomers by knowledge transfer from experienced colleagues, and creating a whistleblowing procedure. Learning to recognize ethical decisions is more challenging than making ethical decisions by computer security professionals and researchers (Fleischmann, 2010).

Regarding AI ethics, according to Forsythe (2001), AI practitioners with a background in computer science or similar IT topics are trained in viewing the world from a technical perspective and undervalue the importance of the social, missing out on the significance of socio-cultural and historical processes as part of the context of use (Bailey & Barley, 2019). The traditional exclusion of ethical analysis from engineering practice is also addressed in Peters et al. 2020), where going beyond the basic ethical demands concerning safety, security, and functionality is rare, leading to the (feared) unintended negative consequences of AI regarding justice, bias, and other societal harms. As a cause, the composition of Agile teams from programmers, designers, and managers hinders the continuous reflective process of ethical impact evaluation and the inclusion of stakeholders beyond the functional needs. As Spiekermann and Winkler (2020) summarize, ethics *by* design (what Dignum calls ethics *in* design) is not ordinary in the IT industry. Their paper describes 14 process requirements and 20 recommendations to guide this approach. Empowerment of the engineers to engage with ethical issues supported by a change of work culture are central. Thus, these authors propose *value-based engineering* as a methodology based on their experiences working with the IEEE P7000 standardization process, which, amongst others, resulted in the IEEE Ethically Aligned Design document (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019).

As Rességuier and Rodrigues (2020) put it, AI ethics needs to have the teeth of ethics and not be limited to the generalist or deontological view of ethics as a law. Hagendorff (2020) argues for virtue ethics aiming at values and encouraging practitioners to aspects such as autonomy, self-responsibility, and broadening the scope of action. In short, such studies see ethics as a way to navigate the murky waters of the use context. For this, the above-mentioned ethical compass or *ethos* is vital.

3.3 In-action ethics and tensions

Ethical guidelines tend to overlook the implications of highly-experienced AI practitioners working for many years. For highly-experienced AI practitioners working on AI projects, a set of ethical guidelines may not always speak directly to them. As the practice becomes routine, knowledge tends to become more automatic and tacit (Schön, 1983), and the practitioner may reflect less on what (s)he is doing. Through reflection, one can make explicit, tacit understandings and create a new sense of the familiar situation differently. This reflection can become immensely relevant when unexpected and new complex issues arise, as may be the case with AI design.

Drawing on Schön (1983), Frauenberger et al. (2017) propose the "In-Action Ethics" framework in the field of Human-Computer Interaction, which accounts for the wide range of real-world situations and contexts in which people use technology and where designers are active human stakeholders (Frauenberger et al., 2017). An important observation by Schön (1983) is the difference between Reflection-in-action and reflection-on-action. The former implies reflection while practitioners are in the midst of an event. On the other hand, the latter involves thinking back after something has happened. Reflection-in-action is used by Frauenberger et al. (2017) to argue that designers' ethics are tacitly in action.

We choose The In-Action Ethics framework as a conceptual lens to examine how ethical issues unfold in AI practice, as recounted by practitioners in the interviews. We select this framework because it views ethics and design as inseparable activities. It "calls for ethical processes to be responsive to issues as they arise in the design, inclusive of stakeholders and reflective as an activity" (Frauenberger et al., 2017, p. 234). The authors argue that ethics awareness needs to be pervasive in the whole design process and make the responsibility shared among the different involved stakeholders. They suggest a set of concepts to operationalize the In-Action Ethics framework. More specifically, they propose the concept of *ethos building*, with *ethos* being understood as "an embodied and intrinsic set of moral positions that tacitly guide actions and decisions" (Frauenberger et al., 2017, p. 234). Moreover, enacting ethos means "doing the right thing", and it is a guiding principle built and maintained by using and reflecting on it (Frauenberger et al., 2017).

Following Varela, we understand that ethical expertise does not stem from rules or reasoning but from the skilled behavior in which people engage daily (e.g., working, talking, moving). Thus, an ethical expert is "nothing more or less than a full participant in a community". (Varela, 1999, p. 24; Kember, 2003, p. 11). As Varela, we attempt to move away from seeing ethics as reasoning to instead focus on actions that stem from immediate coping with a given situation. With this in mind, in this study, we use the

term enacting when referring to putting ethics into practice since it denotes how a subject of perception "creatively matches its actions to the requirements of its situation". (Protevi, 2006, p. 169).

In our study, we also differentiate among individual (or personal), project (or team), and organization (or institution) ethos. There may be situations where enacting these ethe are misaligned (i.e., an individual ethos may not be the same and even conflict with a project or organization ethos). Thus, we define such situations as 'ethos tensions', drawing from the definition of ethical tensions (Bushby et al., 2015). These authors describe ethical tensions as events in the professional practice that raise morally troubling concerns that involve ethical uncertainty, ethical distress, or ethical dilemmas. Ethical uncertainty arises when individuals are uncertain if a particular situation is a moral problem or which moral principle they should apply. Ethical distress occurs when individuals know the right thing to do (enacting ethos) but feel inhibited to act due to organizational regulations, resource constraints, and legal matters, among other reasons. Lastly, *ethical dilemmas* happen when individuals confront equally pleasant or unpleasant mutually exclusive situations (Bushby et al., 2015, p. 212). These three indicators, ethical uncertainty, ethical distress, and ethical dilemmas, have guided the identification of ethos tensions in the interview transcripts.

There are other kinds of tensions that may occur while designing AI systems. For instance, Tatar argues that "design exists because of the tension between what *is* and what *ought to be*". (Tatar, 2007, p. 415). In the author's view, design tensions point at a constrained resource or choice among criteria rather than a problem or solution. These tensions may find the configurations that enable or disrupt a system. Another example is Whittlestone et al. (2019). They point out the necessity to look at tensions to bridge the gap between principles and practice and affirm that it is an essential process "towards an ethics of AI that is practical and action-guiding". (Whittlestone et al., 2019, p. 197). With tension, they refer to "any conflict, whether apparent, contingent or fundamental, between important values or goals, where it appears necessary to give up one in order to realize the other". (Whittlestone et al., 2019, p. 197).

Putting one principle into action requires a previous interpretation of that principle, which may differ by individual organizations (Smit et al., 2020) or people working in them. With more interpretation required, tensions will probably emerge. As suggested by a previous IS literature review on AI guidelines (Smit et al., 2020), the emerging tensions open the door to investigating the practitioners' interpretation of how they apply ethical principles in their practices. Therefore, our study was not exclusively limited to ethos tensions but included other ethical tensions in AI design.

4 Research approach and methods

4.1 Data collection

We conducted thirteen semi-structured interviews in two rounds in English with AI practitioners via Zoom due to the Covid-19 pandemic. The interviews aimed to investigate AI practitioners' perceptions and awareness of AI ethics issues in their past and present work designing AI systems. We were also interested in how the participants handle ethical issues in their everyday practice and consider the different stakeholder groups their systems will impact.

In the first interview round, we used the EU "Ethics guidelines for Trustworthy AI" (AI-HLEG, 2019) as common ground and a prompt to investigate how responsible AI is considered when designing AI systems for the public sector. The link to the Guidelines was included in the email invitation for the interview to provide a concrete, accepted, standard frame of reference to engage with a delicate and abstract topic such as ethics. Using snowball sampling, we recruited practitioners working for the Swedish public sector who 1) have long experience working with AI and 2) work with AI in the public sector. All the interviewees were located in Sweden and had Swedish as a working language. Seven were men and two women in the first round; the second round consisted of four follow-up interviews with a selection of these interviewees, all male. The interviews were conducted from October 2020 to December 2021 and lasted 40 minutes on average. Participants were recruited through purposive or snowball sampling from participants who recommended other interviewees.

The first round of interviews was structured in four parts. First, the interviewees answered questions regarding their background and their current jobs. Second, we asked them questions about the specific AI technologies they were developing, such as providing examples of AI systems, the expected users of these technologies, and how they strive to make them ethical. Third, we asked them to rank the key requirements in the Guidelines in order of importance for their work; and whether they thought their co-workers would rank them differently. Fourth, they answered several questions we formulated, building on the pilot version of the Trustworthy AI Assessment List (included in the Guidelines) to analyze how they integrated ethics into their practice. As described in section 2.1, we focused on three specific principles (transparency, fairness, and stakeholder consideration and involvement). The second round of interviews aimed to elucidate why such tensions occur and how they are experienced in AI design practice. We combined the findings of the second round of interviews with those of the first round to create a holistic picture.

Interview participants

Most of the study participants are highly experienced in working with AI, with over ten years of experience on average. Their backgrounds range from computer science, IT, mathematics, and AI to economics. Their roles vary considerably, but many of them hold senior positions and primarily work as developers of AI artifacts. Some worked for the same organizations, and some co-worked on the same AI projects. Examples of AI applications they were developing or leading are decision support, risk assessment, document summarization, image classification, speech recognition, dialogue systems development, and machine translation. Most of these tools were developed for internal usage by the organizations' employees.

Their job responsibilities differed depending on their role, and most interviewees have worked on diverse and multiple projects for the Swedish public sector. Some of them were not employed by a government agency but by a State-owned research institute that closely collaborates with Swedish government agencies. Appendix 1 summarizes the interviewees' roles, educational background, duties, and years of experience with AI.

4.2 Data analysis

The interviews were recorded, fully transcribed (using smooth verbatim), anonymized, and qualitatively analyzed by coding the interview transcripts. We used MAXQDA software to work systematically with the coding of the transcripts. The coding work was performed iteratively and collaboratively within the research team.

We started with coding inductively as a way to enter the data analysis with an open mind (Saldaña, 2021). We used various coding methods: In Vivo, Process, and Descriptive coding. We categorized the resulting 624 codes into seven overall themes that reflected the main topical areas: 1) Lack of guidelines, 2) Ethics conceptualizations, 3) Transparency, 4) Fairness, 5) Stakeholder consideration and involvement, 6) Risk or impact assessment and 7) Accountability.

After the first coding round, building on the In-Action Ethics framework and the literature on tensions (Tatar, 2007; Bushby et al., 2015; Whittlestone et al., 2019), we identified some ethos and other tensions in the data. In the second coding round, we used Versus coding, which identifies dichotomous conflicts in groups, individuals, organizations, or processes and is appropriate for studies of conflict and opposing norms and value systems (Saldaña, 2021). Three main themes emerged from this second coding round: 1) tensions in interpreting ethical principles, 2) tensions in enacting ethical

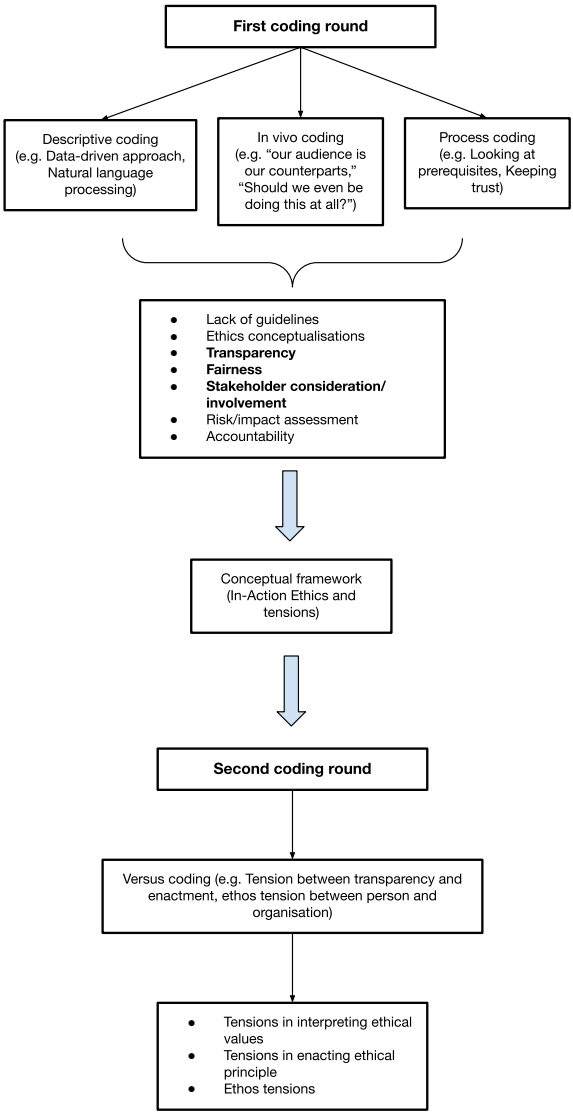

Figure 1. Data analysis process

principles, and 3) ethos tensions. The data analysis process is depicted in Figure 1. The findings of the data analysis are discussed in the following section.

5 Findings and analysis

5.1 Tensions in respondents' ranking of the ethical principles

To prompt thinking about ethics in their work, we revisited the seven key requirements from the Guidelines. We showed the list of requirements and asked the participants to rank them in order of importance in their work designing AI systems for the public sector. One of the nine interviewees preferred not to rank the requirements. Table 1 summarizes how many respondents ranked a specific requirement as one of the top three most important principles (i.e., they ranked it as either 1, 2, or 3).

Privacy and data governance, and transparency were ranked the highest requirements. Curiously, technical robustness and safety was one of the lowest-ranked requirements. However, respondent R1 said that it is "almost a given" coming from an engineering background, which may explain it.

Key requirement	*Number of respondents that ranked it in the top three*
Human agency and oversight	3
Technical robustness and safety	2
Privacy and data governance	5
Transparency	5
Diversity, non-discrimination and fairness	4
Environmental and societal well-being	3
Accountability	2

Table 1. Respondents' ranking of the seven key requirements of the EU "Ethics Guidelines for Trustworthy AI". (AI-HLEG, 2019)

This ranking gives a snapshot of how different AI practitioners view and consider the requirements differently. By no means does it provide a whole picture of how they reason

around the use of such ethical areas. However, it is interesting that there is no consensus on the relative importance of ethical principles. In the following subsections, we unpacked the participants' interpretations of the three selected ethical principles (*transparency*, *non-discrimination and fairness*, and *environmental and societal well-being*).

5.2 Tensions in the interpretation of ethical principles

Throughout the interviews, we typically found multiple ways particular ethical principles can be construed. By asking participants how a specific principle was applied in their AI projects, we uncovered the multiplicity of meanings and their polysemic understanding of ethical principles at work. For example, an interviewee acknowledged such diversity in interpretations for transparency, noting that depending on how one relates to it may lead to a different answer.

> **Transparency means a lot of different things for me**. If we speak about transparency towards the citizens and the public in general, which I'm guessing that you are most interested in, I would say that that kind of transparency is **quite low today**. We have some information on our website, and we advertise, or we make aware that we are using AI as a part of our development of things like that. But, **on a more detailed level, no, we're not quite transparent**. (R4).

AI transparency to the public is "quite low today" for this respondent due to not being transparent on a "detailed level". From the quote, it is hard to know which detailed information was R4 referring to. A common conceptualization of transparency within the AI community is algorithmic transparency, meaning that transparency equals giving explanations of how a system works. For instance, R5 thought that since transparency is hard to apply in an AI environment, explanations are necessary:

> It's a little bit difficult [to apply transparency] when it comes to public administrations, as we need the traceability. It's hard to prove that in an AI environment, especially in deep neural networks, **it's hard to explain exactly what is happening**. And, since we have that **requirement on us today to be able to explain how we come up with a decision, transparency in a wider perspective is really, really important**. (…) We need to **explain as much as we can**, and make it open for everyone I would say, to keep the trust". (R5)

Explanations are not limited to the model itself (the "black box" model as referring to deep neural networks) but also about data and the "feeling of control" R5 mentioned.

5.3 Tensions in enacting ethical principles

We asked the participants how they applied ethical principles (such as transparency or fairness) in their design practices. We found that enacting principles is not straightforward but requires an existing interpretation of the principle at hand and how to approach it in practice, which is usually open to multiple interpretations. On top of that, tangible constraints will limit how principles can be considered and reflected in the work practice (e.g., regulations, resources, incentives, unclarity).

In regards to transparency, none of the interviewees mentioned any specific tool, methodology, or measure used to assess or implement transparency in the designed AI models. This lack of formalized assessment created tensions by not knowing how to enact transparency in practice. A respondent who worked with developing chatbots described what they are currently doing to implement transparency: they receive user feedback:

> I don't think **we have gotten that far to have a methodology for implementing transparency**. (...) If you look at the chatbot, it's pretty transparent to ask the question and receive the answer. And you can judge to see if it's the answer that fits the question or not, so you can have a mechanism for **feedback from the user** to see if this is a good answer or not, but that's so far how we got with that. (R6)

R6 acknowledges that a specific methodology to implement transparency is lacking in the organization where they work. Their current approach is to check whether the chatbot provided suitable answers through the users' feedback. However, confusion between transparency and accuracy may occur (i.e., the chatbot seems accurate rather than transparent by proving the correct answers).

The analysis also reveals that interviewees usually point at bias measurement when asking how fairness is ensured. Bias is a highly discussed topic in the AI community, and most participants mentioned that it is an issue they are trying to tackle in their teams. Generally speaking, there was an awareness that bias can be introduced in many ways, such as through the developers or the training data. However, bias is not something that can be easily "solved," as this interviewee points out:

> So, what we can do is actually **highlight to people that we're working with that**, this is, this particular language model is trained on this **particular data**, and you will see biases in there, and you will have to be **aware that there will be biases** and we **cannot claim to solve the biases** and the bias problems automatically, but the people should be **aware of them**. (R1)

The respondent claims that "highlighting to people that they are working on fairness" is a way to work towards fairness. People's awareness is critical to handling bias.

When ranking the key requirements, several participants shared their concerns about the environmental impacts of their models, which in a sense is an ethical and moral consideration per se vis-à-vis the impact of AI on the planet and human existence. We found tensions between the participants' awareness of working with AI models consuming energy and global natural resources and the participants' concerns about not finding ways to assess the environmental impact of AI models (specifically, large language models) that some participants were working on. This concern is something that was (at the time of the interview) being discussed within their teams:

> We **are having discussions internally in our group about the environmental effects** of training large language models over and over again, **so I would say that is probably my top three issues here because it's like a hidden thing that you don't see as a researcher**. You kick off the training of staff, and it goes for… you run it for several days, and it consumes lots of energy, and you don't see where that energy is coming from, or what is doing to the planet in a sense. (R1)

R1 thought that environmental issues were some of the most pressing issues in AI design since it is somehow hidden and hard to evaluate. This causes worry to them, and it is currently being discussed within their teams. The key requirement in the Guidelines on *Societal and environmental well-being* emphasizes that the AI system's design should be assessed to make it the most environmentally friendly way possible. However, according to the participants, this seems to be hard to implement.

Tensions between governance and professional practice.

We also found certain occasions in which governance may clash with the AI design practice, for example, regarding transparency. Several respondents expressed the difficulty of being transparent with AI design beyond the technical aspects due to organ-

izations' internal rules and legal constraints. Regarding the latter, an interviewee who works as a product owner mentioned:

> Currently **we are not very transparent** on which areas we are using AI, and we also, if we go down to a very detailed level and we look at the models…. **We have some legislation** that will actually say that we cannot reveal the specifics of that kind of model. For sure, we could discuss it in a more general term, like we do in this meeting, or like I also discussed in these kind [sic] of models on a general level with other public agencies, how we work in general with methods, **but the specifics we would keep fairly tight**. (R4)

It seems as if the respondent would like to be more transparent (i.e., revealing the model's specificities), but the law is constraining them. Another participant who, between the first and second round of interviews started working in the private sector noticed that being transparent while being a researcher working for the public sector is very different than in the private one, primarily because of the business secrecy, admittedly not related to the AI system at hand:

> I had felt some discomfort in not being able to be transparent, but that is mostly because **I transitioned from being a researcher where we are really transparent to more of a business perspective where we don't want to spill all our beans to every client. There's some secret sauce.** (R1)

The striving for fairness also caused tension between the legal requirements and value application. When speaking about fairness, some interviews addressed the issue of using protected features. Due to the Swedish Constitution, some protected features (e.g., political views, gender, or sexual orientation) cannot be used in the AI models. However, by proxy, the AI models would find correlations among such protected features, leading to discriminative impacts (Barocas & Selbst, 2016). In recent years, such an issue has received much attention in the machine learning community, with conferences like ACM FAccT having fairness issues as their primary goals. A respondent shared their tension and discontent with how this issue was handled. They were aware of techniques and methods to increase fairness; however, they were not working with them. Instead, they remove such protected attributes before using the model:

> So, by proxy, we're still using gender, **but we don't see it**. It's **not that transparent**, and there are other techniques and methods that actually do to increase the

> fairness, but that's how we currently are more or less working with **taking out data before we model to ensure fairness**. (R4)

5.4 Ethos tensions

Regarding the broader societal impacts of AI, most interviewees mentioned that this is not assessed as part of their job. An interviewee thought there is not enough discussion within the governmental agencies about the societal transformations driven by AI implemented on a broader scale. This respondent's tension was epitomized when they said that the question "Should we even be doing this [AI design] at all?" should be asked much more and earlier than what is asked now:

> I would say that the first question I would ask, and I think this is a question that is not often asked when we talk with others about AI and transparency, accountability, and things like that, is the question, "**Should we even be doing this at all?**" (...) and what actually caught my interest when talking with other government agencies is **how little thought is going into** "should we actually be doing this at all". That question, either it has been passed over, or that question is already answered in some way. (...) So I would say that the **"should we" question should be asked a lot more and a lot earlier**, and I think those are the interesting questions. (R4)

This quote illustrates the tension between personal and organizational ethos. On the one hand, the respondent's question is if organizations are responsible enough to do AI design. On the other hand, the respondent enquires about the responsibility behind deciding that AI is the way to go. Focusing on the impacts and effects of an AI system, organizations may ignore the actual need and responsibility to design AI systems.

Ethos tensions may also emerge among fellow workers. The following quote was shared when, in the follow-up interviews, we asked for a situation in which their ethos conflicted with someone else's. The participant emphasized the importance of how things are done, while their co-worker was more focused on the result. They shared their thoughts this way:

> I think that in my opinion, I think that how we do things matters. **So how we approach this problem actually matters**. (…) S**o, I will fight for my opinion**. I remember one discussion more explicit than others: we were discussing how

> we do things matter, and this other person that I was having the discussion with was saying that 'but yes, will reach the same result in the end.' And I will say, yes, we do reach the same result in the end, **but how we do things matters**, how we **discuss** these things will matter, how we **express ourselves**, how we **actually discuss,** and what **kind of culture we want to have within [organization]**. And the other person we're [sic] not seeing in the same way, because I think he was more focused on the end result being the same. (R4)

The respondent showed a strong opinion towards the importance of how one gets to their final aim. Indeed, the respondent did not believe that the end justifies the means and firmly believed that the modus operandi influences the culture within one's work organization.

6 Discussion

This study aimed to investigate how AI practitioners perceive ethical principles in their work regarding AI design and how they interpret and enact them in practice. From the analysis of the interviews with AI practitioners working in public organizations, we learned that designing AI systems brings tensions in several aspects of the practice. Such tensions emerge when interpreting ethical principles from the top-down guidelines and enacting them in professional practice.

Firstly, ranking ethical principles allowed us to see that practitioners interpret and consider ethical principles related to AI design differently. The ethical principles in Responsible AI ethics frameworks can be interpreted in myriad ways, and practical guidance on how to do so is usually not offered (Schiff et al., 2021). Moreover, for highly experienced practitioners (as the participants in this study), AI ethics frameworks may not always speak directly to them. As the practice becomes routine, knowledge becomes tacit, and consequently, reflection about one's practice may be less habitual (Schön, 1983). However, such reflection becomes crucial when new complex issues arise. Reflecting during the course of an event (reflection-in-action) shows that ethics are tacitly in action (Frauenberger et al., 2017). In other words, "ethics is not an appendage to design but an integral part of it" (Devon & van de Poel, 2004, p. 461). Thus, we argue that reflection and discussion on ethical considerations and values should be given a more central space in the practitioners' day-to-day work so ethical reflection can be embedded into AI practitioners' practices.

Secondly, we found tensions in the different practitioners' interpretations of ethical principles. Ethical principles such as transparency and fairness have many interpre-

tations, not necessarily all corresponding with how AI ethics guidelines define them. Some of this variety is already acknowledged by the participants, especially in the case of transparency, but this is not always the case, as in fairness. There are over twenty notions of fairness, but it seems as if individual fairness is commonplace among the interviewees. We argue that the more "technical" definitions of algorithmic transparency and fairness, which do not consider the broader sociotechnical context, are too limiting (Veale et al., 2018; Dignum, 2019; Schiff et al., 2021). In fact, we argue that algorithmic systems are not just technical objects made of "code and data but an assemblage of human and nonhuman actors" (Ananny & Crawford, 2018, p. 983). Thus, "opening the black box" may not be sufficient to ensure transparency. Similarly, algorithmic fairness needs to consider the broader social context for better and fairer systems (Selbst et al., 2019).

Thirdly, tensions also exist in enacting those AI ethics principles in practice. Reasons behind this include the lack of formalized processes to assess such principles and constraints with governance. Moreover, a few participants shared their concerns about the environmental issues of training AI models, which are ethical issues per se. Striving to apply some principles might be misaligned with legal constraints or organizations' internal rules. Along with Morley et al. (2020) and Vakkuri et al. (2020), we also believe that the lack of suitable resources and tools limits the integration of ethical principles into AI design practices. As Seppälä et al. (2021) and Mayer et al. (2021) pointed out, the implementation of AI ethics is still in a formative stage even though there is widespread awareness of AI ethics guidelines. More interpretation is required with higher abstractions to express ethical principles (Smit et al., 2020). With more interpretation required, tensions emerge. This opens the door to investigating the practitioners' interpretation of how they apply ethical principles in their practices. Therefore, identifying tensions is vital when ethical principles are meant to be operationalized, as Smit et al. (2020) stress.

Finally, we identified ethos tensions between personal ethos and organizational ethos. Rakova et al. (2020) remind us of the essentiality of considering AI practitioners, organizational structure, and human culture when studying AI's social impacts. Designing AI systems is an iterative process, with "different people at the table, different information flows, different normative relationships, different authority structures, and different social and environmental considerations in mind" in each stage (Devon & van de Poel, 2004, p. 461). Thus, if AI design is framed as an inherently social practice, ethos tensions are to be expected.

6.1 Contributions

The study contributes to the AI ethics field by providing a clearer understanding of how AI practitioners perceive ethical issues related to their day-to-day work. Contrary to the studies which focus on understanding if AI ethics guidelines are being implemented (or not), we focused on the practitioners' perception and enactment of ethical principles in broader terms. Conducting interviews instead of surveys helped us better understand the humans behind designing, developing, managing, and assessing AI systems.

Taking the lens provided by the In-Action Ethics framework (Frauenberger et al., 2017) and previous work on tensions (Tatar, 2007; Bushby et al., 2015; Whittlestone et al., 2019), this study suggests that tensions enable understanding the multiple meanings that can be associated with ethical terms often used in the public discourse of AI ethics. Furthermore, discussing tensions regularly within the project or organization remains key as ethical considerations are context-sensitive and may change over time. In this vein, we argue that the identified tensions may function as a lever for change, not as a sign of a lack of ethical competence in public organizations. We claim that understanding the different tensions that can occur in practice and how they are dealt with daily are crucial to contributing to an empirically-grounded Responsible AI. Responsible AI ultimately aims to develop AI technologies that enhance societal and environmental well-being (Dignum, 2019).

How practice deviates from principles reflects a fruitful adaptation to the circumstances. To study the perception and enactment of AI practitioners is to understand how this mediation principle-practice takes place. Studies such as this contribute to IS literature by emphasizing that ethical frameworks do not reflect how principles are put into practice (Smit et al., 2020; Jantunen et al., 2021; Mayer et al., 2021; Seppälä et al., 2021). On the contrary, researching the complexities of enacting ethical principles reveals that tensions can occur, and how such tensions are handled will affect the eventual design of the technology.

The In-Action Ethics framework calls for AI ethical assessments to be adaptable in response to the ethical tensions that emerge during the design. AI practitioners should be encouraged to be reflexive all over the AI systems' lifecycle through, for instance, ethical mentoring (Waycott & Vines, 2019) or regular team discussions on ethical issues that come up during work practice. This could improve AI practice by making it more open to discussing ethics and encouraging reflection and discussion on the different ways to tackle emerging tensions.

Some of the tensions described in this study are not exclusively found in public sector institutions. For instance, Moss and Metcalf's (2020) interviews with and observations of "ethics owners" in Silicon Valley practitioners also found some of the described

ethos tensions when trying to operationalize ethics (Moss & Metcalf, 2020). However, in our study, some interviewees' tensions related to transparency could be most prominent in governmental agencies due to organizations' internal rules and legal constraints. Sweden has a long tradition of government transparency based on the right-of-access principle (offentlighetsprincipen), which gives citizens the right to access information about the state's and municipalities' activities[3]. Moreover, most interviewees ranked transparency as one of the top requirements in the Guidelines, along with privacy and data governance.

This study expands recent IS literature on operationalizing AI ethics (e.g., Smit et al., 2020; Wang et al., 2020; Jantunen et al., 2021; Mayer et al., 2021; Seppälä et al., 2021) by providing another perspective on the human dimension in AI design. We argue that focusing on ethical tensions that emerge throughout AI design is an essential step toward studying AI practice in vivo and in situ. Identifying tensions, their source, and how to deal with them contribute to understanding the experiences and challenges that AI practitioners find themselves in. This is a unique characteristic of this study that the other IS studies we found to date did not incorporate.

The ethical tensions described in this study also contribute to IS design research. The overall methodology of the AI Ethics Maturity Model described by Jantunen et al. (2021) is based on design science research. Specifically, we argue that studying the tensions that may emerge while putting the ethical requirements into play will help organizations reach higher maturity levels. Optimized, the last of all maturity stages describes maturity levels of a proactive approach where ethics are indeed considered. Identifying and solving emerging tensions could help ensure the Optimized maturity level is reached. Future research may even connect the maturity levels to the kinds and severity of the ethos tensions observed in an organization. This implies that the maturity level of an organization may be contingent on the system under development.

6.2 Limitations and future research

The small number of participants, the sampling procedure (i.e., snowball or purposive), and the fact that the vast majority of participants are not native English speakers and men are limitations of the study. Moreover, our participants are located in Sweden and work in or for a specific number of public organizations. Finally, the prompts used (the Guidelines) may configure the discussion. Nonetheless, the participants showed diversity in thinking, which led us to consider that the Guidelines may not have impacted that much.

Understanding the study participants' perceptions in practice led to two main findings that future studies should further explore. First, it is essential to find helpful instruments, procedures, and methods to assess ethical principles enactment in practice. Second, tensions should be more deeply investigated and even be the subject of discussions in internal meetings within projects or organizations. Discussions on the diverse interpretations of ethical principles and how such principles are (or should be) enacted would contribute to a richer understanding of how AI systems can be designed and aligned with human values and social contexts.

Akin to Tatar's argument, ethical tensions conceptualize AI design "not as a problem solving but as a goal balancing". (Tatar, 2007, p. 415). Tensions allowed us to step back from technical design and situate AI systems as sociotechnical. Since AI systems are deployed in the real world, it is evident that tensions among perspectives will exist. Therefore, we should take such tensions as a starting point for discussion among diverse stakeholder groups.

Frauenberger et al. (2017) encourage organizations to support a working culture that gives space and structure for ethos building and care, for instance, via ethos building workshops. Moreover, having an ethos 'facilitator' within a project, someone responsible for promoting ethos building would be a further step toward operationalizing ethics in action. In such workshops or similar 'Ethical AI debate clubs', tensions could be further discussed by drawing on real or fictional cases to generate questions while encouraging reflection-in-action. This is analogous to the recommendation by Spiekermann and Winkler (2020) to enroll an interdisciplinary value expert. The ranking exercise and developing additional methods and tools to discuss ethos would be valuable teamwork activities to share perceptions and experiences around how these principles are connected and applied in work practice.

Finally, it is worth further investigating the implications of AI design in the public sector. For instance, the Swedish government has declared AI for the public sector to be highly prioritized. Studying how this sector differs from private and non-profit sectors would also be relevant. In this study, we selected governmental agencies as public sector organizations. Other extensions would include other public organizations such as local government agencies or areas such as academia or healthcare since they all have their own challenges and agendas.

Notes

1. We use the term enaction, enacting, or enactment when referring to putting ethics into practice. We base it on Varela's conception of "cognition as enaction," as *enaction* connotes

grounding activity in concrete actions and engaging with reality by taking actions (Varela, 1999).

2. The government agencies in Sweden are state-controlled organizations that act independently to carry out the policies of the Government of Sweden.
3. For more information, see https://www.regeringen.se/sa-styrs-sverige/grundlagar-och-demokratiskt-deltagande/offentlighetsprincipen/.

Acknowledgments

The authors thank the participating AI practitioners for their time and for sharing their thoughts, views, and experiences. We also thank the guest editors and anonymous reviewers for their valuable comments and suggestions.

References

Ada Lovelace Institute, AI Now Institute, and Open Government Partnership. (2021). *Algorithmic accountability for the public sector*. https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

AI-HLEG. (2019). *Ethics Guidelines for Trustworthy AI*. European Commission https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

Albrechtslund, A. (2007). Ethics and technology design. *Ethics and Information Technology*, *9*(1), 63-72. https://doi.org/10.1007/s10676-006-9129-8

AlgorithmWatch. (2020). *Automating Society Report 2020*. AlgorithmWatch GmbH. https://automatingsociety.algorithmwatch.org/

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973-989. https://doi.org/10.1177/1461444816676645

Bailey, D. E., and Barley, S. R. (2020). Beyond design and use: How scholars should study intelligent technologies. *Information and Organization*, *30*(2), 100286. https://doi.org/10.1016/j.infoandorg.2019.100286

Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671-732. https://dx.doi.org/10.2139/ssrn.2477899

Bednar, K., Spiekermann, S., and Langheinrich, M. (2019). Engineering Privacy by Design: Are engineers ready to live up to the challenge? *The Information Society*, *35*(3), 122-142. https://doi.org/10.1080/01972243.2019.1583296

Bellman, R. (1978). *An introduction to artificial intelligence: can computers think?* Boyd & Fraser Publication Company, San Francisco.

Bushby, K., Chan, J., Druif, S., Ho, K., & Kinsella, E. A. (2015). Ethical tensions in occupational therapy practice: A scoping review. *British Journal of Occupational Therapy*, *78*(4), 212-221. https://doi.org/10.1177/0308022614564770

Bynum, T. W. (2008). Milestones in the history of information and computer ethics. In K. E. Himma & H. T. Tavani (Eds.), *The Handbook of Information and Computer Ethics* (pp. 25-48). John Wiley & Sons, Inc, Hoboken, New Jersey.

Cave, S. (2020). The problem with intelligence: its value-laden history and the future of AI. In A. Markham, J. Powels, T. Walsh & A.L. Washington (Chairs), *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)* (pp. 29-35). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3375627.3375813

Crawford, K., and Calo, R. (2016). There is a blind spot in AI research. *Nature*, *538*(7625), 311-313. https://doi.org/10.1038/538311a

De Vries, K. (2020). AI policy in the Netherlands: More focus on practice than principles when it comes to trustworthiness. In S. Larsson, C. Ingram Bogusz & J. Andersson Schwarz (Eds.), *Human-Centred AI in the EU: Trustworthiness as a strategic priority in the European Member States* (pp.132-157). European Liberal Forum asbl, Brussels.

Devon, R., and van de Poel, I. (2004). Design ethics: The social ethics paradigm. *International Journal of Engineering Education*, *20*(3), 461-469.

Dignum, V. (2019*). Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer Nature Switzerland, Cham.

Fleischmann, K. R. (2010). Preaching What We Practice: Teaching Ethical Decision-Making to Computer Security Professionals. In R. Sion, R. Curtmola, S. Dietrich, A. Kiayias, J. M. Miret, K. Sako, & F. Sebé (Eds.), *Financial Cryptography and Data Security* (pp. 197-202). Springer, Berlin, Heidelberg.

Forsythe, D. (2001). *Studying those who study us: An anthropologist in the world of artificial intelligence*. Stanford University Press, Stanford.

Frauenberger, C., Rauhala, M., & Fitzpatrick, G. (2017). In-Action Ethics. *Interacting with Computers*, *29*(2), 220-236. https://doi.org/10.1093/iwc/iww024

Government Offices of Sweden. (2019). National approach to artificial intelligence. Swedish Ministry of Enterprise and Innovation. https://www.government.se/4a7451/contentassets/fe2ba005fb49433587574c513a837fac/national-approach-to-artificial-intelligence.pdf

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

Hopkins, A., and Booth, S. (2021). Machine learning practices outside big tech: How resource constraints challenge responsible development. In M. Fourcade, B. Kuipers, S. Lazar & D. Mulligan (Chairs), *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)* (pp. 134-145). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3461702.3462527

Jantunen, M., Halme, E., Vakkuri, V., Kemell, K. K., Rousi, R., Mikkonen, T., Duc, A. N., and Abrahamsson, P. (2021). Building a Maturity Model for Developing Ethically Aligned AI Systems. In B. A. Farshchian (Ed.), *IRIS 2021: Papers of the 44th Information Systems Research Seminar in Scandinavia* (Article 5). IRIS Association. IRIS. https://aisel.aisnet.org/iris2021/5

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Kember, S. (2003). *Cyberfeminism and artificial life*. Routledge, London.

Mayer, A. S., Haimerl, A., Strich, F., and Fiedler, M. (2021). How corporations encourage the implementation of AI ethics. In F. Rowe, R. El Amrani & M. Limayem (Chairs), *European Conference on Information Systems (ECIS 2021).* Association for Information Systems, Atlanta, Georgia, USA. https://aisel.aisnet.org/ecis2021_rp/27

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501-507. https://doi.org/10.1038/s42256-019-0114-4

Morley, J., Floridi, L., Kinsey, L., and Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, 26(4), 2141-2168. https://doi.org/10.1007/s11948-019-00165-5

Moss, E., and Metcalf, J. (2020). *Ethics Owners: A New Model of Organizational Responsibility in Data-Driven Technology Companies*, Data & Society Research Institute. https://apo.org.au/node/308440

Perera, H., Hussain, W., Whittle, J., Nurwidyantoro, A., Mougouei, D., Shams, R. A., and Oliver, G. (2020). A study on the prevalence of human values in software engineering publications, 2015 -- 2018. In G. Rothermel & D-H. Bae (Chairs), *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering (ICSE '20)* (pp. 409-420). Association for Computing Machinery, New York, NY, USA.. https://doi.org/10.1145/3377811.3380393

Peters, D., Vold, K., Robinson, D., and Calvo, R. A. (2020). Responsible AI—two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, *1*(1), 34-47. https://doi.org/10.1109/TTS.2020.2974991.

Prior, M., Rogerson, S., and Fairweather, B. (2002). The ethical attitudes of information systems professionals: outcomes of an initial survey. *Telematics and Informatics*, *19*(1), 21-36. https://doi.org/10.1016/S0736-5853(00)00014-9

Protevi, J. (2006). *A dictionary of continental philosophy*. Yale University Press, New Haven.

Rakova, B., Yang, J., Cramer, H., and Chowdhury, R. (2021). Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. In J. Nichols (Ed.), *Proceedings of the ACM on Human-Computer Interaction*, *5(CSCW1)* (pp. 1-23). Association for Computing Machinery, New York, NY, USA.. https://doi.org/10.1145/3449081

Rességuier, A., and Rodrigues, R. (2020). *AI ethics should not remain toothless!* A call to bring back the teeth of ethics. *Big Data & Society*, *7*(2), 2053951720942541. https://doi.org/10.1177/2053951720942541

Russell, S. J., and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson Education Limited, Essex.

Saldaña, J. (2021). *The coding manual for qualitative researchers*. SAGE Publishing, Thousand Oaks.

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., and Lennon, M. (2021). Explaining the Principles to Practices Gap in AI. *IEEE Technology and Society Magazine*, *40*(2), 81-94. https://doi.org/10.1109/MTS.2021.3056286

Schön, D. A. (1983). *The reflective practitioner: how professionals think in action*. Basic Books, New York.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. In d. boyd & J. Morgenstern (Chairs), *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* '19). Association for Computing Machinery, New York, NY, USA, 59-68. https://doi.org/10.1145/3287560.3287598

Seppälä, A., Birkstedt, T., and Mäntymäki, M. (2021). From Ethical AI Principles to Governed AI. In J. Valacich, A. Barua & R. Wright (Chairs), *Conference on Information Systems (ICIS 2021)*. Association for Information Systems, Atlanta, Georgia, USA. https://aisel.aisnet.org/icis2021/ai_business/ai_business/10

Smit, K., Zoet, M., and van Meerten, J. (2020). A Review of AI Principles in Practice. In D. Vogel, K. Ning Shen & P. Shan Ling (Chairs), *Pacific Asia Conference on Information Systems (PACIS 2021)*. Association for Information Systems, Atlanta, Georgia, USA https://aisel.aisnet.org/pacis2020/198

Spiekermann, S., and Winkler, T. (2020). Value-based Engineering for Ethics by Design. *SSRN Electronic Journal.* https://dx.doi.org/10.2139/ssrn.3598911

Spiekermann, S., Korunovska, J., and Langheinrich, M. (2019). Inside the organization: Why privacy and security engineering is a challenge for engineers. *Proceedings of the IEEE*, *107*(3), 600-615.. https://doi.org/10.1109/JPROC.2018.2866769

Susser, D. (2019). Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures. In V. Conitzer, G. Hadfield & S. Vallor (Chairs), *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)* (pp. 403-408). Association for Computing Machinery, New York, NY, USA.. https://doi.org/10.1145/3306618.3314286

Tatar, D. (2007). The design tensions framework. *Human-Computer Interaction*, *22*(4), 413-451. https://doi.org/10.1080/07370020701638814

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). Ethically Aligned Design: A vision for prioritizing human well-being with autonomous and intelligent systems. IEEE. https://ethicsinaction.ieee.org/wp-content/uploads/ead1e.pdf

Vakkuri, V., Kemell, K. K., Jantunen, M., and Abrahamsson, P. (2020). "This is just a prototype": How ethics are ignored in software startup-like environments. In V. Stray, R. Hoda, M. Paasivaara, & P. Kruchten (Eds.), *Agile Processes in Software Engineering and Extreme Programming (XP 2020)* (pp. 195-210). Lecture Notes in Business Information Processing, *383*. Springer, Cham. https://doi.org/10.1007/978-3-030-49392-9_13

Varela, F. J. (1999). *Ethical know-how: Action, wisdom, and cognition*. Stanford University Press, Stanford.

Veale, M., Van Kleek, M., and Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In R. Mandryk & M. Hancock (Chairs), *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)* (pp. 1-14). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3173574.3174014

Verbeek, P. P. (2006). Materializing morality: Design Ethics and Technological Mediation. *Science, Technology, & Human Values*, *31*(3), 361-380. https://doi.org/10.1177/0162243905285847

Wang, Y., Xiong, M., and Olya, H. G. T. (2020). Toward an Understanding of Responsible Artificial Intelligence Practices. In T. Bui (Ed.), *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICCS 2020)* (pp. 4962-4971). HICSS Conference Office, USA. https://doi.org/10.24251/hicss.2020.610

Waycott, J., and Vines, J. (2019). Designing technologies with older adults: Ethical tensions and opportunities. In B. Barbosa Neves & F. Vetere (Eds.), *Ageing and Digital Technology* (pp. 173-187). Springer, Singapore.

Whittlestone, J., Nyrup, R., Alexandrova, A., and Cave, S. (2019). The role and limits of principles in AI ethics: towards a focus on tensions. In V. Conitzer, G. Hadfield & S. Vallor (Chairs), *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)* (pp. 195-200). Association for Computing Machinery, New York, NY, USA.. https://doi.org/10.1145/3306618.3314289

Winfield, A. F., Michael, K., Pitt, J., and Evers, V. (2019). Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems [Scanning the Issue]. *Proceedings of the IEEE*, *107*(3), 509-517. https://doi.org/10.1109/JPROC.2019.2900622

Appendix 1. Summary of the participants' background

	Educational background	*Role*	*Org*	*Duties/Responsibilities*	*AI experience*
R1	Computational linguistics and software engineering	Senior researcher	A	Developing AI tools for the Swedish public sector, project leader, technical implementation, and contact point with customers.	>20
R2	Mathematics and computer science	Senior researcher	A	Developing AI tools for the Swedish public sector, including strategic planning, working on methods, bias assessment.	>20
R3	Computational linguistics and philosophy	Senior researcher	A	Developing AI tools for the Swedish public sector, including strategic planning, working on methods, bias assessment.	>20
R4	Economics and IT	Product owner	B	Developing AI solutions and services to create business value in a Swedish governmental agency.	10
R5	Business and IT	Senior advisor	C	Advising the Government of Sweden on how to use AI in the public administration.	>20
R6	Mechanical engineering and IT	IT Strategist	B	Developing IT strategy for a Swedish governmental agency.	4
R7	Computer science and knowledge management	Senior advisor of the IT unit	D	Project Manager for diverse AI projects of a Swedish governmental agency and AI advisor, AI developer, and maintainer.	5
R8	Natural science and IT	Project leader	D	Project manager for an AI project in a Swedish governmental agency.	2

R9	Machine learning and computational linguistics	Research director	E	Leading R&D strategy, product definition, and features, product research in a private company.	16
R10	Cognitive science and computer science	AI ethics researcher	F	Researching the social impact of algorithms and data, testing algorithms for bias and fairness.	13
R11	Design and philosophy	Senior researcher	A	Developing ethical frameworks.	14

Table 2. Summary of the interviewee's background, roles, organizations, duties, and years of experience working with AI. The interviewees in the *italic* style participated in the second round.