Trust, Artificial Intelligence and Software Practitioners: an interdisciplinary agenda

Sarah Pink[1], Emma Quilty[1], Rashina Hoda[2] and John Grundy[2]
Emerging Technologies Lab[1], Human-centric Software Engineering Lab[2]
Monash University, Melbourne, Australia
sarah.pink@monash.edu, Emma.Quilty@monash.edu, rashina.hoda@monash.edu,
john.grundy@monash.edu

Introduction

Trust and trustworthiness are central concepts in contemporary discussions about Artificial Intelligence (AI)-based software and the relationships between people and AI, with an expansive body of literature and debate relating to human trust in AI and automated systems and technologies (Shahrdar et al, 2019). There is also significant existing research regarding how concepts of trust and trustworthiness are used in industry and policy domains (Khan et al, 2023). However, as existing critiques reveal (Pink 2022, 2023, Reinhardt 2022, Schneiderman 2020) both concepts have been defined and used differently across software engineering and computer science disciplines (Shneiderman, 2020; Lu et al, 2022; Ahmed et al 2023), in agendas of industry and government, in the humanities and in the qualitative social sciences (Pink 2022; Reinhardt 2022). Thus, while agendas for trustworthy AI superficially engage shared terminology, they fact frequently represent diverse and contested understandings of what trust, trustworthiness and ethics entail, and different alignments with the relations of power and capital.

Recent research suggests that in industry narratives trust is often treated as a commodity or resource (Pink 2023), whereby there has been a "discursive commodification of trust" (Krüger & Wilson 2022). The notion of "trustworthy" AI frequently refers to certain ethical principles, which are thought to be able to win peoples' trust (Pink 2022). Yet, software Engineers have identified serious limitations with existing methods for designing ethical, trustworthy AI-based systems (Ahmed et al 2023; Lu et al 2022) and assumptions that AI technologies and systems can be trusted, or that trust can be gained, won and mobilised have been contested by scholars, usually on the basis that they fail to account for the characteristics of human trust (Freiman 2022, Pink 2023). These concerns, from both software engineering and anthropology, reveal a major gap between the ambitions of industry and government in achieving trust in technologies on the one hand and the everyday realities of how trust is constituted and experienced on the other.

This situation suggests that the question of trust in AI, and the development of trustworthy AI both remain wicked problems, which require a new vision and approach (Schneiderman, 2020; Lu et al 2022). In this article we suggest that to begin to approach these problems we need to both attend to questions of how trust comes about in sociotechnical relations, and to add a new element to the analysis - a consideration of the work practices, experiences and conceptualisations of trust of the software practitioners working in and interfacing with the software industry who are tasked with creating trustworthy AI. Existing anthropological research has revealed that the subjectivities of tech workers can be integral to the algorithmic technologies they generate (e.g. Seaver 2022, Lanzeni & Ardevol 2017), and this point is equally applicable to AI. However, little is known about how software engineers who design and develop AI-based software applications conceptualise trust in AI as they go about their everyday work practices or about the implications of this for sociotechnical relationships (Khan, 2023; Ahmed et al 2023).

In this article we take a first step towards engaging with these questions through an interdisciplinary approach. Whereas existing approaches have tended to draw on psychology, philosophy and economics (Andras et al 2018), we take a novel approach which brings together sociotechnical perspectives in software engineering with design anthropology. In doing so, we first map out the key approaches and concerns of these disciplines with respect to trust and trustworthiness. We then report on how we examined our common concerns through a study of how trust and trustworthiness are articulated and performed by software practitioners. We subsequently ask what the implications of our findings are for two pressing questions: what steps are needed to advance a shared interdisciplinary understanding of trust which encompasses industry, practitioner and everyday human engagement with AI; and on what basis might trust be better integrated in the training and skills of future software practitioners.

Trust, Trustworthiness and AI

There is an expanding literature surrounding trust and trustworthiness with reference to AI. However, to date these discussions tend to be siloed, with software engineers advancing technical standards, philosophers and ethicists[1] debating abstract concepts, while anthropologists have explored the status of trust and ethics through empirical ethnographic investigation. We next outline this discussion from the perspectives of sociotechnical software engineering and design anthropology respectively.

Sociotechnical software engineering perspectives

A growing body of research and practical techniques have been proposed by AI experts and software engineers to address emerging problematic issues in the engineering of AI-based software systems. Glikson and Wolley (2020) characterise how AI differs from other technologies and review 20 years of empirical studies into the development of "trust" in AI-based systems. They describe how the manifestation of AI e.g. as robot, virtual, and embedded software, and its level of capabilities constitute key antecedents to the development of trust. They propose a framework that describes how various AI characteristics shape users' cognitive and emotional trust. They identify how the tangibility, transparency, reliability and immediacy of behaviour of AI impacts the development of cognitive trust, and the nature of anthropomorphism develops emotional trust. Jacovi et al (2021) describe an attempt at formalising the concept of trust in AI-based systems. They examine the nature of trust in AI, what are prerequisites to developing trust in AI, and propose a model of trust inspired by the concept of interpersonal trust. Toreini et al (2020) examine the relationship between trust in AI and the machine learning technologies they are built upon. They claim that in order to build AI-based systems that users and stakeholders can trust, the trustworthiness of the machine learning components of the systems must be deeply understood. Their systematic approach to building trustworthy machine learning components builds on the ABI social science framework and considers fairness, explainability, auditability and safety of machine learning technologies.

A number of theories, processes, techniques and tools have also been developed in recent years to address the need to ensure AI-based software is engineered in ethical ways and can thus meet trust requirements of users. Khan et al (2022) survey a range of works relating to ethics of AI

[1] Our focus is on trust and trustworthiness, rather than on the extensive debate on AI ethics, we engage with questions related to ethics here where they overlap explicitly.

from a software engineering perspective. They identify a global convergence onto 22 ethical principles, including transparency, privacy, accountability and fairness. A lack of ethical knowledge and vague principles are identified as major challenges to ensuring ethical AI (Pant et al. 2023). Ahmad et al (2023) systematically review a large number of empirical studies proposing requirements engineering techniques for AI-based systems i.e. approaches to determining what such systems should do, not how they do it. They identify that many traditional approaches to determining software requirements must be rethought when developing AI-based software. Buhnova et al (2023) examine recent progress in research on trust in software engineering across various application domains. They make the point that many systems are moving towards co-operating autonomous agents, requiring a rethink of concepts of mutual trust. They identify the need for new social metrics of trust, governance mechanisms needed, and trust assessment.

Moreover, in the software engineering field, many approaches to improving the trustworthiness of AI-based systems have been proposed in recent years. Hutchinson et al (2021) review practices around machine learning datasets. They identify that the creation of datasets that underlie many machine learning approaches and hence AI-based systems lack visibility into their creation processes. They make a case for greater transparency about data, accountability for decisions made in dataset creation to ensure more trustworthy AI-based software. Arnold et al (2019) propose a framework for declaring the conformity of AI-based systems to guidelines and regulations. Lu et al (2022) propose a roadmap for ensuring responsible AI via the use of appropriate software engineering methods. Shneiderman (2021) proposes a set of guidelines to ensure more responsible AI-based systems engineering. Vakkuri et al (2022) review a number of guidelines proposed to date and note the lack of impact on industrial practice. They note the need for approaches that address the novel requirements for software development of AI-based systems.

Perspectives from philosophy, legal studies and design anthropology

Recent reviews (Pink 2022, 2023, Reinhardt 2022) have discussed the rise of the concept of trustworthiness in discourses about AI, and the quest to design trustworthy AI. "Trustworthy AI" initiatives abound and are advanced by consultancies, such as Deloitte, industry organisations such as IBM, and the European Commission.[2] Indeed, as commented elsewhere, "The concept of *trust* has come to be associated with an anticipated future in which Automated Decision Making (ADM) and Artificial Intelligence (AI) will have been successfully, ethically, inclusively and responsibly implemented in ways that solve societal problems and increase efficiency, safety and quality of life" (Pink 2022), and many large technology companies are involved in designing and testing a range of AI technologies for user and public trust.

However, as Reinhardt argues from the perspective of practical philosophy, "currently, AI ethics tends to overload the notion of trustworthiness" (Reinhardt 2022). Reinhardt's analysis suggests that trustworthiness "runs the risk of becoming a buzzword that cannot be

[2] See for example : https://www2.deloitte.com/be/en/pages/strategy-operations/strategy-analytics-mergers-acquisitions/solutions/trustworthy-ai.html, https://research.ibm.com/topics/trustworthy-ai, and https://ai-watch.ec.europa.eu/topics/trustworthy-ai_en.

operationalized into a working concept for AI research". They identify several serious deficiencies in existing assumptions in one of the most significant documents referring to ethics and trustworthy AI: Europe's High-Level Expert Group's *Ethics guidelines for trustworthy AI,* ethics guidelines. They argue that "despite an apparent consensus that AI should be trustworthy, it is less clear what trust and trustworthiness entail in the field of AI and what ethical standards, technical requirements and practices are needed for the realization of TAI." They reveal a set of conceptual and practical inconsistencies in this framework in such a way that suggests that it is very unlikely that the vision of trustworthy AI it proposes will ever come about (Reinhardt 2022).

Indeed, a wider set of critiques of the notion of trustworthy AI have been developed. For example, following a philosophical definition of trustworthiness, Brusseau (2023) argues that it is not possible that AI could be trustworthy. Indeed, a wider set of critiques of the notion of trustworthy AI have been developed. For example, Freiman argues that trustworthy AI is "conceptual nonsense" Freiman (2022). They outline a series of definitions of trust developed by philosophers and ethicists and dissect a number of existing critiques of the notion of trustworthy AI. These suggest that trust between humans and machines is impossible and therefore argue that the only possible objects of trust are either the institutions or people responsible for technology. Some legal studies scholars have followed a similar argument to suggest there is "little sense to speak of trust in AI devices" or "to regulate risks for trust as a way to solve this problem" (Estella 2023, no page numbers), as well as to caution against "horizontal regulatory law" without "without further sectoral regulation or standardization" (Laux et al 2023: 25). Freiman concludes that "Ultimately, the scholarly fields of trust and AI ethics are about power, social justice, and activism" and suggests that scholars in these fields should become activists in order to remain relevant. Freiman's critiques are well founded, but we propose a different conclusion; scholars of trust and AI ethics do need to shift the ground of analysis but rather than simply as activists arguing from the sidelines, they should do so as practical scholars, contributing within interdisciplinary teams. We elaborate on this later, in this section we next outline the role of empirical research in advancing alternative definitions of AI trust and ethics.

One of the limitations of the critiques of trustworthy AI that are advanced by humanities scholars, such as legal studies scholars, philosophers and ethicists, is that the models of trust discussed are infrequently actually witnessed in everyday life (Freiman 2022). For example legal scholars focus on the "public": calling for future policy to focus on "public values, trustworthy public institutions, and trustworthy private technological infrastructures" (Bodó, & Janssen, 2021: 17); suggesting a increase in individuals's insight into AI decision making models, public trust and subsequently greater ease "for the public authority to obtain the cooperation and obedience of the public" Williams (2022: 480); and that "engagement between the user and the model" is needed for human and societal trust in AI" (Hacker et al 2020). While these are important points they do not account for how these trust relations will be created by people in everyday life.

Much empirical knowledge about trust used in disciplines such as philosophy, as well as by industry and government organisations is gathered in quantitative survey research. This means it relies on reported information, derived from questions people answer about trust. These are often taken at face value to indicate that if people trust (or would trust) a person, institution of machine or not, rather than on the realities of how, when and where trust comes about and is experienced in everyday life. Thus, we also need to also interrogate and account for trust as an experienced reality, which is manifested and accessible through fieldwork encounters between

ethnographers and participants in research, rather than exclusively through statistical means. As philosopher O'Neill has pointed out, the statistics delivered by polls about public trust do not tell us anything about how people make judgements about whether "to trust or refuse trust to particular individuals or institutions for particular matters" and do not account for the care that people may take in doing so (O'Neill 2017). A design anthropological perspective accounts for such detail and does so by going beyond suggesting that trust can be present or absent in a relationship, interaction or transaction between humans, or humans and technology. Instead, design anthropology turns around the relationship between trust and those other agents that might be related to it. It begins with the premise that if we understand trust as an underpinning element (and experiential possibility) of the circumstances of human life, we can then ask how AI and other technologies and things might become part of those circumstances of trust (Pink 2023).

To address some of these issues, design anthropological approaches to trust regarding self-driving cars and digital data have been developed using ethnographic research findings (Pink 2022, 2023). Design anthropologists and anthropologists of ethics understand everyday life as processual, "excessive, uncertain, and emergent" (Mattingly & Throop 2018: 482), and as a site where people continuously improvise to cope with contingent circumstances, where change might be visible or barely perceptible (Smith & Otto 2016, Pink 2023). Ethics thus emerge from particular circumstances and are contingent, and rather than being fixed or unchanging, in everyday life are adapted "to suit situations (and the limitations of these situations) as they unfold" (Pels 2001, Pink 2017, 2022). Trust has similarly been seen as emerging in everyday situations and is, like ethics, changeable and contingent on the specific circumstances people find themselves in. This approach to trust is critical to the idea that a person would trust a technology. However, it takes the critique in a different direction in suggesting that trust should not be treated as a transactional relationship (Corsin 2011: 178) but rather defines trust as "a *feeling*, or category of feeling, which describes [a particular kind of] anticipatory sensations" (Pink 2021) regarding what will happen next. Existing ethnographic research (e.g. Pink, Lanzeni and Horst 2018; Pink, Osz, Fors & Lanzeni 2021) demonstrates that "to trust" something does not necessarily involve a decision that an interaction or transaction with that AI technology, person or organisation will be trustworthy. Rather, it is "a sensory experience of feeling or disposition towards something" whereby trust is seen as a mode of "confidence based on familiarity" (Frederickson 2016: 59) related to "how we feel about what might be going to happen next", or a "feeling between what we know and what we think we know" (Pink 2021). Thus neither AI, other people or organisations can be inherently trustworthy nor ethical, as trustworthiness and ethics are likewise contingent and circumstantial, rather than fixed qualities that can reside in unchanging or permanent ways in AI systems and technologies, people or organisations. In this understanding likewise trust is not a tangible, fixed entity that can be commodified, won or gained (Pink 2023, Krüger & Wilson 2022). Therefore the dominant definitions of these concepts are difficult to mobilise when it comes to real everyday encounters with AI. Indeed, it is therefore difficult to either design or develop trustworthy and ethical AI, and it is similarly difficult for people to encounter AI technologies as trustworthy or ethical in everyday life. This leaves an unanswered question relating to how trustworthiness and trust appear in the ordinary work of software engineers building AI-based systems, and what the implications of this are for AI and ethics and education.

Interdisciplinary issues

Across the software engineering, sociotechnical and social sciences there is growing interest in and interrogation of the questions of trust and trustworthiness in AI. These disciplines approach trust and trustworthiness from different directions - engineers by seeking to make the

technology trustworthy, philosophers by defining trust as an abstraction, and anthropologists by investigating trust in everyday life. Across these disciplines a range of different, and not always directly aligned approaches to the concept of trust have been mobilised, to variously seek to understand the interactional, transactional, circumstantial, contingent and experiential aspects of trust and its development.These different approaches and their commentaries demonstrate that the question how trust and trustworthiness for AI can be achieved has not yet been resolved, and that further interdisciplinary sociotechnical research that explores the interwoven nature of social and technical aspects (Hoda, 2021) is needed. One issue that has been identified as ripe for further analysis, and that requires such an interdisciplinary approach, is the question of the role of software engineers of AI-based systems who play a key role in the practical making of what is hoped to be trusted and trustworthy AI, which we elaborate further in the next section.

Building AI-based systems

The making of AI-based technology has become a common everyday task, carried out by developers in a range of different organisational spaces (Khan et al 2023, Lu et al 2022). Many studies have been undertaken by software engineering researchers into practices of building software systems in general and AI-based systems in particular. Studies have looked at what ethical concerns software engineers have and their own concepts of trust in relation to AI-based software, and to what degree software engineers feel they can address these ethical concerns (Widder et al, 2023). Several studies have looked at the challenges and approaches to identifying requirements for AI-based systems i.e. what they should do, not how they should do it (Belani et al 2019; Ahmed et al 2023). Developing better techniques for engineering complex AI-based software has been a very active recent research area (Ebert et al, 2023 ; Amershi et al 2019). Concepts of ethical AI, responsible AI, trustworthy AI and approaches to reason about these have been popular (Schneiderman 2022; Lu et al 2022). Several recent studies have identified that software engineers lack adequate training and education around the development of AI-based systems, including the engineering of ethical and responsible AI-based software (Heck et al 2021; Bublin et al 2021; Kastner et al 2020). Many researchers have identified that engineering trustworthy AI-based systems is a little-understood challenge that is already taking place but lacks adequate processes, tools and understanding by practitioners (Schneiderman 2022; Ahmed et al 2023; Lu et al 2022). Increasingly software engineers themselves utilise AI-based tools to build AI-based systems (Carleton et al 2020). This brings an interesting challenge that software engineers need to trust their own AI-based work supporting tools, often not fully understanding what the tools do and how they do it (Weisz et al 2022). Several very recent studies have investigated implications of using generative AI-based tools to engineer software systems (Ozkaya, 2023). It is as yet not fully understood how such tools help - or hinder - the development of trustworthy AI-based software. The degree of trust in these AI-based software engineering tools by software engineers is as yet unclear, and yet they are already being widely used.

Anthropologists have emphasised the subjective and contingent circumstances through which such technical work is carried out. Seaver (2022) analyses the work of the makers of music recommendations, to suggest that they navigate perceived oppositions between algorithms and humans on the one hand, and computing and taste on the other. Seaver engages with anthropological and sociological theories of taste, as a means of encountering the relationship between human subjectivity and the apparent rationality of algorithms and technology. In doing so, he urges us to consider how technology and taste cannot be separated in the making of algorithmic systems, and seeks to analyse this through his ethnographic encounters with engineers. Lanzeni and Ardevol (2017) have also emphasised the subjective modes through

which technology future imaginaries are constituted and performed by tech workers. Their studies correspond with existing research about how trust is experienced by tech workers. For example, Pink et al (2018) in research with people who worked in tech design, blockchain and makerspace environments found that while they had to depend on them in their work, they didn't see digital data or technologies as trustworthy or reliable. The recurring pattern in such studies is thus the urge to account for how human feelings, imaginaires, aspirations or anxieties become part of the work of software engineers and other tech industry workers. Seaver (2022) seeks to account for taste, whereas Lanzeni and Ardevol (2017) interrogate how futures are imagined and Pink et al (2018) examine anxiety and trust. We take a similar approach to ask how trust is articulated in the work software practitioners do, what their understandings of trust are, how these subjectivities are performed in their work practices, and what the implications of these findings might be for training future software engineers and other tech industry workers.

There has been a focus on trust in teamwork in many fields. Unsurprisingly research in organisation studies and psychology suggests that trust relations support successful work in teams (Costa et al 2018: 180). This growing field of study has overlapped with recent research about software developer teams. For instance, Amirali Sajadi and colleagues (2023) identify conflicting definitions of trust in this field, and highlight the need to define trust and how it appears in the contexts of software development. Their study, which analyses trust in the context of interpersonal relations between developers based on their interactions on github, employs a scale of dimensions of trust, using psychology and organisational behaviour theories, with an overarching aim of being able to automatically identify and quantify the instances of different dimensions of trust to computationally evaluate levels of trust in teams. This work reveals that defined as such dimensions of trust can be identified in teamwork amongst software developers, and offers a relevant background for understanding how some of our participants experienced the dynamics of teamwork. However, our own approach differs in two ways: our analysis of trust is qualitative, it engages with and seeks to draw out the experiences and commentaries of research participants, to understand how trust is articulated and manifested in their work practices; and rather than theorising trust as interpersonal, we engage a design anthropological approach to theorising trust as contextual and emerging from broader sets of contingent circumstances beyond interpersonal interactions.

Research design and methods

Our ethnographic research investigated how trust mattered and was experienced and articulated by practitioners in the software industry. We sought to understand how on the one hand this aligned with the tendency to commodify trust in dominant narratives, and the debates regarding how and if AI can be trustworthy, which we have outlined in the previous sections. In doing so we were also interested in defining where software practitioners believed trust and trustworthiness should or could be encountered.

Our interdisciplinary approach leveraged the expertise and experiences of SE4AI researchers and of anthropologists of AI and trust to collectively script an ethnographic interview outline. Ethnographic interviews (O'Reilly 2011) are semi structured and conversational, leaving scope for the researcher to follow threads of discussion which can enable in depth responses to questions and are flexibly adapted to each participant's experience, rather than relying on a pre-existing one-size-fits-all script. The interviews covered how research participants' definitions and practical and affective experiences of AI, trust and trustworthiness as well as how they might design for trust or seek to generate trust in client/user engagement, with a total of nine software practitioners developing AI-based software.

Anthropological research depends on deep engagements with participants, rather than massive research sample. For instance anthropological studies have been based on long interviews with as few as one or two participants (e.g. Shostak 2000, Desjarlais 2003). To understand how trust is perceived in the process of software development we recruited nine participants to represent a broad group of professionals in the industry, rather than focusing solely on software developers/engineers. Our participants were: two applied engineers (chemical and mechanical engineers) to investigate how automated software is used; two marketing and strategy managers to understand the process of delivering and communicating software systems; one software team manager to gain an understanding of the role trust plays within a software development team and between the team and clients; and four software engineers. Table 1 lists the participants roles, the pseudonym they were assigned in order to report individual responses and protect their identities, and an outline of the key systems they work with/build, processes they use and approaches to trust requirements specification and implementation (where shared with us). Participants were recruited through existing contacts in the team in the respective software engineering and computer science fields and through snowball sampling where participants identified other potential participants in their own networks.

Our participants worked in several diverse industries using AI-based technologies, including control systems (Ahmed, Jason, Catherine, Kei, Mia and Charlie); CRM systems (Ella, Mia); Internet of Things (IoT), vision and other sensor-based systems (Ahmed, Jason, Kei, Mia); Timetabling systems (Oliver), military simulation systems (Oliver, Charline); mining and water treatment systems (Jason, Catherine); and art, branding and promotional systems (Mia). The types of AI components in these systems also varied considerably, ranging from movement and light detection (Ahmed); data analysis (Ella, Henry, Charlie); natural language processing (NLP) and text mining (Henry, Mia); intelligent controllers (Ahmed, Jason, Catherine, Kei, Charline); robotics (Oliver); human in the loop decision making (Catherine, Jason, Mia, Charline); and simulation systems (Oliver, Mia, Charlie). Some told us about development processes used to build these AI components. These included agile methods (Ahmed), issue resolution-driven continuous integration (Ahmed, Ella, Oliver), data collection, wrangling, curation and integration processes (Ella, Oliver, Mia, Charline); various model training approaches including using realistic scenarios and human feedback (Ella, Henry, Jason, Catherine, Mia); rapid prototyping (Henry, Oliver); and obfuscation of data and/or creation of synthetic data, often to protect privacy (Mia, Charlie). We asked about specific ways trust-related issues were specified and refined into target software system AI-based models. Approaches included privacy policies and security policies (Ahmed, Ella, Oliver, Catherine, Mia); identifying and testing for potential unbalanced data and/or biassed models (Mia, Charline); human-in-the-loop over-rides (Jason, Catherine, Charlie); certification requirements (Ahmed, Jason); end user feedback to refine requirements after deployment of models (Ella, Ahmed, Henry, Jason, Catherine); and testing developed models on realistic scenarios to refine requirements, implementation and training (Ella, Jason, Catherine, Mia, Charlie).

Because our participants were situated differently in the industry, this gave us an opportunity to analyse how their approaches to trust cohered or differed. However, we emphasise that it was not our objective to compare how different participants from different industries defined trust in relation to the different technical requirements specifications their work involved. Instead, the aim of our interdisciplinary approach was to draw on an anthropological perspective to analyse their approaches to trust from the starting point of their experiences.

Pseudonym	Job Title	Software Systems/ AI Technologies
Ahmed	Software Developer	Software control systems for IOT devices; room occupancy detection, light etc auto-adjustment
Ella	Lead Engineer	Customer relationship management (CRM); data analysis, work automation
Henry	Team Lead	Recommender systems; Natural Language Processing (NLP), text mining, semantic reasoning
Oliver	Product Manager	Timetabling, Military simulations; simulation and optimisation
Jason	Mechanical Engineer	Mining systems; Image Detection, robotic control, sensing systems
Catherine	Process Engineer	Water treatment systems; Supervisory Control and Data Acquisition (SCADA) automation, monitoring, feedback
Kei	Strategy & Development	Internet of Things (IoT); control, sensing
Mia	Marketing & Brand Manager	Art creation, brand promotions; Internet of Things (IoT) control, sensing
Charlie	Software engineer	Scheduling software, education software, health research software; data mining and analysis, recommenders

Table 1. The interview participants (using pseudonyms for confidentiality)

Anthropological ethnographers conventionally undertake fieldwork with participants in the sites of their everyday life. In the case of our research, undertaken during the Covid-19 pandemic both we and our research participants worked from home, and our fieldwork was undertaken online. Recent literature about online ethnography, largely augmented by the pandemic, has demonstrated that it is possible to generate in depth ethnographic materials by engaging with participants through video calls and conversations (Pink 2021, Howless 2021). Due to the COVID-19 lockdowns in Australia our research was also undertaken online. Participants were interviewed, for an average of one hour, via Zoom giving us access both to their homes where they were working, and to the platforms they used through screen-sharing, which as Ritter (2021) also notes can offer modes of investigating their ways of knowing and practice. Screen-sharing allowed participants to take us through the specific tools they used, and to use diagrams when they needed to illustrate a particular point or story.

The interviews were transcribed verbatim and the full transcripts were imported into NVivo qualitative analysis software. The data was initially coded in NVivo using the categories of interview questions including: the affective dimensions of trust, AI and automation, designing for trust, trust and trustworthiness and clients and user engagement. The next stage of analysis involved writing up a summary of the key findings, organised using the interview categories as well as longer ethnographic vignettes. This allowed the data to be analysed dynamically and iteratively, moving between the findings, authors and theoretical conceptualisations. Whilst NVivo is useful as a software analysis tool as it makes the process more efficient and allows one to explore different pathways in the data, ultimately there is ‘no mechanistic substitute for the complex processes of reading and interpretation’ (Hammersley & Atkinson 2007, 167).

Findings

In this section we outline the key findings from analysis of our interviews regarding how participants understood AI, trust and trustworthiness.

Understandings of AI

Acknowledging that uses and applications of AI are always contextual and that definitions of AI are contested across stakeholder groups, rather than impose a definition of AI on participants, we asked participants themselves how they define AI in general and in their work. While at a more fine grained level participants tended to cite their own experiences and sometimes different views regarding how AI would become part of their work and society, most participants concurred that intelligence or smartness as a key defining characteristic of AI. They differentiated AI from other modes of automation or automated decision-making (ADM) in the sense that AI can learn and AI-powered machines can be "taught". For example, one participant described how their company overcame the issue of facial recognition through the use of synthetic data, to "teach" the machine to detect "people" rather than recognising them. Others described AI as adding in technologies like sensors, for example, to "give the software eyes, and then they can program it to make decisions … getting towards the state where they can almost fully autonomise the system" (Oliver, Product Developer). Participants also spoke from their everyday life experiences to define AI, for instance, one invoked the example of a kettle, suggesting "maybe it monitors your phone, it learns what time you wake up and what time you have coffee every day, and it automatically turns itself on. (Henry, Software Engineer Team Lead).

Two software engineers we interviewed, pointed out that AI is over-hyped to the point people believe it can do things that in reality it cannot. One reduced it to simply being a set of techniques, which they said "could be used to predict something, that can be used to help automate something, that could be used to assist users". Another participant described the AI "magic box" as simple "smoke and mirrors", especially at the beginning of the process. They suggested that hype might be when marketing teams or project leads over-sell the possibilities of AI to clients; or when developers get carried away with doing something "cool" with AI, and shifting away from the core purpose of problem-solving for end users. Participants had differing views regarding whether humans should be involved when AI was used or implemented, but they noted that AI often requires a lot of tweaking and adjusting, especially at the beginning of the process, and that AI is expensive to build and this can be a lengthy process.

Understandings of trust

Trust is a complex concept, because it is used with different meanings and nuances across different academic disciplines, government and industry contexts, and in everyday life, usually in ways that are often not clearly defined. Participants did not cite theories of trust derived from academic or grey literature but presented what sociologists see as "common sense" ways of thinking about trust, sometimes reflecting on everyday life experiences to emphasise how it was part of their lives. For example:

> Society is based on sharing trust. When you go to a restaurant you put very, very, very big trust in the person who is serving your food, the person who cooked your food. It's so easy to make you sick or just add something in your food and you may never notice it because they make it so tasty with all the spices. But we trust them. (Ahmed, Software Developer)

Another participant described his preference for instant work communication to come via Slack rather than Facebook messenger:

> I guess cause probably not very logical reasons I guess because I started using it when it was pretty small I've been using it while it progressed into a bigger company but I knew of it when it was smaller and it seems it still has the right values (Charlie, Software Engineer)

Such lay definitions of trust involve people thinking through their lives and looking for ways to exemplify trust, and thus participants tended to draw extensively on their own personal everyday life and work experiences to consider trust. However, simultaneously we found that participants were eclectic in their commitments to models of or views on trust, meaning that there is no single coherent view across our sample, not one single place in the process of developing, selling or using software for automated technology that trust was solely related to. Moreover, strikingly participants did not necessarily explicitly align their ideas about and commitments regarding trust to dominant narratives or initiatives in software engineering towards trustworthy AI, showing a common gap between scientific research and professional practice.

The majority of participants described trust as something that could be gained or given, and generally saw it as something that it was desirable to obtain either from end users, clients or colleagues. However the specific relationships and sites within which they discussed encountering, experiencing or gaining trust varied. For example, Oliver, a product manager at a software company, described the role trust plays in the relationship between himself and clients, focusing on his "discussions that go ahead with clients in the background to try and evoke that, to communicate trust". In some cases participants explicitly said that they trusted automated technologies. For instance Ella, a Lead Software Engineer, told us that "We give trust to the automation to make decisions for us". While Jason, a mechanical engineer, who led a team in the mining industry, told us that he trusts the automated software used in the underground mine he works in because he is able to see the accuracy of the automation from his vantage point as he manages the team. However, he said that some members of his team, especially those belonging to the 'older generation of coal miners', trusted the technology less, sometimes taking over the machine manually. As he put it:

> We had this automation set up a couple of years ago. We tell it where the pitch is and just trust that the machine's going to move itself the right way and not run into something, and that was really hard for especially the older generation of coal miners. (Jason, Mechanical Engineer)

Indeed, it was not only older workers who drew the line between when an automated system might be trusted or not, but also the question might be contingent on the nature of the specific tasks that a machine was required to undertake. For instance, Ella, a Lead Engineer said that in her experience the AI products her team developed could not be used "to deliver critical systems", explaining that "in a medical setting, or like aviation, so all those critical systems that you can't embrace risks".

Yet while the idea that people would trust a technology to undertake a task in practice, the question of where and how trust relationships might start can also be located earlier in the development process. Oliver, the Product Manager, told us when asked about the place of trust in the process of developing a software application, that it is not the system that builds trust. Rather, for Oliver the key question related to how software companies elicit trust from their users is: "It's not that I want a system that builds trust, it's how do you get people to trust something that is being built".

Other participants referred to trust as part of a relationship in work teams, whereby a team leader would trust their team to deliver, as Ella, Lead Software Engineer, expressed it: "We encourage trust where you give developers trust, we trust them to deliver what they think they can". Two participants who work closely with software developers - in marketing - also situated trust as being invested in their wider team, in that they said that they trust that the software they sell has been thoroughly tested before going out to the end user.

Almost all of the participants saw trust is something that builds, accumulates or is created over time. For example, Catherine, a process engineer trained in chemical engineering, told us about how the decision to introduce an automated system used to run the water treatment plant where she works as a process engineer came about (Catherine, Process Engineer). Part of her role involved monitoring the system. She estimated that it had taken two years to convince the board of the plant to automate the system. Concerns about cybersecurity for example, needed to be assuaged. She suggested that

> The only reason we got it by was by spending the money to get this AI system in. By not having to build a new part of the plant, it was going to save them millions of dollars. And [have] all the assurances with cyber security. That was like a huge big deal for them. (Catherine, Process Engineer)

Jason described how it was taking time for the older generation of employees in his company to become familiar with the automated software, often overriding the systems and running the machinery themselves. He described how it was a:

> Classic situation, "I know how to drive this thing. I've been driving it for 15 years. I'm really good at it", but the computer's better. Every single day, the computer's better than them, but they don't trust it. So, they'll do what they think is right, which 90% of the time is, but then that 10% of the time, they're actually wrong, stuff something up, and it causes us a big delay when we break something or bend it the wrong way or those types of things. (Jason, Mechanical Engineer)

For them - he explained - it is hard for them to trust what they cannot see, they cannot see the results since they are stored in the computer, for them the accuracy of their own experience was superior to the automated software. Jason trusted the software more than his team did however - like Charlie - he watched the software closely for glitches, feeding information back to the software developer to improve the system. He trusted the software but often felt like he had to babysit it. In these trust was not an experience that was generated by specific characteristics of the system, but rather it came about when people became open to trusting the system to undertake specific tasks.

We learned that participants saw trust as being contingent on the relationship between a person and a technology, rather than inherent in the design of the technology. For instance, one participant framed trust as something that exists between the end user and the system, *not* as something that is *built* into the system itself. The idea that trust in technology is contingent on the social relations that it becomes part of was further illustrated by Henry, a team leader in software engineering, who pointed out, he might trust the software he creates and sends out into the world *but* he doesn't trust how someone else might take up what he has created. As he expressed it:

> Someone comes along and decides that they're going to use my service with some other stuff that I've not considered, right. Now the trust is broken somehow, now I've got to go back and fix it, but nothing has changed at my end, right. I've not changed the code, I've not changed the model, like the AI part, and I've not changed, I'm still the same person that created it. (Henry, Software Engineering Team Lead)

Participants therefore located trust in different aspects of the process of the development and use of software.. They also variously spoke of how different elements may need to configure to constitute trust, including for instance, user experience, performance, predictive capacity, how it's presented, and explainability. When they spoke of trust they often used conventional definitions that imply that trust is transactional or interactional. Their discussions were also based on their contextual experience and on where they were situated in the industry, such as developers or in marketing or in other roles and in relation to the different fields they referred to such as water treatment, health or mining. In anthropological research one of the aims is often to engage with the specifics of everyday life and contextual experiences and seek to understand the patterns that emerge across different contexts (see Table 1). Indeed in, an analysis of their experiences of trust design anthropologically - to ask where and how trust emerged as situational, contingent and shifting as new circumstances came about - it is clear that there were common factors; the relations and experiences of trust our participants referred to did not necessarily correspond with the idea of trust as something that could be gained, won, or fixed in a technology. In common, when asked to consider trust as part of these processes participants identified trust as something that was constituted in particular relationships, which would be generated, but would be different for people who were differently situated (e.g. generationally or in relation to their expertise), could shift and change, and that was contingent. These experiential dimensions of trust are significant in helping us understand how trust was part of the working lives of the software practitioners we worked with, in a practical sense, rather than in a discursive sense.

Significantly our research also reveals how practitioners spoke about the concept and experience of trust in ways that correspond largely with their sociotechnical relationships at work and in their everyday lives, and they did not refer directly to either the engineering literatures or industry narratives regarding trust in AI that we have outlined in the literature review above.

Trustworthiness

While trust appeared as situational and evolving, in our discussions with participants, trustworthiness was understood as a quality of software. However, similarly it was not necessarily seen as a fixed quality or as something that could exist independently of the relationships within which the software was applied. In addition the ways in which participants discussed trustworthiness depended on the specific nature of the roles they played in the industry, as set out below.

The participants who worked directly in the design and development of software such as engineers and developers, described trustworthiness as a dimension of the software itself, as something that can be coded into the software architecture, but that would not necessarily be activated or apparent initially. Two participants suggested, for example, that trustworthiness and people's trust in the software could be measured or evaluated through the process of development and use. For instance, Henry, a software engineering team lead, suggested that at the beginning of the development process there would be "many assumptions of why it [the software] wouldn't be trustworthy" meaning that they would ask themselves: "what do we need

to be able to tweak to get this system working, to then get to the point where evaluating trust makes more sense". Ella, a lead software engineer, likewise suggested that with the "architecture and characteristics for data security, security in all those things" in place, then "we could come with measurement for trustworthiness, and trust itself". The idea of measuring trustworthiness was also expressed by Ahmed (also a software developer) who posed: "Can we establish trustworthiness metrics for headphones, for example, and then…the government could manage and define the metrics… and [provide] tax relief". That is, could the government incentivise products that have software that meets trustworthiness metrics.

Three participants in product and marketing were more inclined to describe trustworthiness as something that can be communicated between a project manager and a client, or a user and a software application. This resonated with their understanding of trust as something that would come about in these same relationships. However in the case of trustworthiness, like the participants working in software development and design they treated trustworthiness as a quality of the software, which nevertheless needed to be made visible before it would engender relations of trust. For example Oliver, who was a project manager suggested the "the soft aspects of [a] system, like trustworthiness" could be communicated on screen media, via an app or web service to explain, as he put it "what is this app doing with your data, and you can download all that data history". He suggested that "high fidelity mock-ups" could be used during the design process to show these "soft aspects".

Summary of findings

In this section we have outlined the findings of our ethnographic study in relation to three key questions: how did participants understand AI and how did they discuss trust and trustworthiness in relation to AI-based software. We found that participants saw AI as a technology with machine learning characteristics, which tended to be overly hyped, while in fact it was already part of everyday life. In relation to trust, participants based their discussions of trust on their own experiences or commonsense logics, rather than on current debates in academic or industry literatures. Generally participants understood trust as a contingent relationship or experience which would evolve over time, which corresponds with definitions of trust developed in design anthropological studies of trust in everyday life reviewed above. However their discussions of trust, and the relationships in which they assumed trust would come about tended to correspond to the roles they played in the software industry. In contrast to trust, participants tended to be more uniform in terms of their understandings of trustworthiness, which all attributed as a quality of software in ways that correspond to some degree with the engineering literatures reviewed above. Yet in their narratives the extent to which trustworthiness of software was activated or became visible to its users depended on the specific development of the relationships within which it was used or demonstrated.

Recommendations for research and practice

From our AI software engineering practitioner participant feedback, a key issue is defining for a system under development (or indeed deployed) what is "trustable AI-based software"? Participants indicated that such a definition is not straightforward and can depend on the nature of AI-based components, nature of the software system, perception of end users and so on. They also reported that the definition of trustworthy AI-based software varies depending on the perspective taken: developers, sellers, deployers, users, government, society. On a related issue, practitioners reiterated the need to obtain buy-in to trust when developing and deploying AI-based software. Trust is like security - it is very hard if not impossible to retrofit trust to developed and deployed software systems if not conceptualised, defined, designed-in early on.

Several participants noted the user and system relationship that is quite different with AI-based software to traditional software systems and the need to achieve and maintain a trust relationship between user and system. Participants noted the need to gain this trust buy-in at the user, organisational and governance levels is critical to acceptance of AI-based software.

As noted above our research was designed to investigate perceptions and approaches to trust, rather than to compare how participants understood trust in the context of specific technical requirements specifications, AI-based software design, implementation and testing. In so doing we wanted to be able to offer a more abstract analysis of the ways that trust is articulated amongst people who work in this industry. We regard this as the first analytical step that needs to be taken in order to view the way trust comes about from a fresh interdisciplinary perspective. However, we acknowledge the importance of capturing and monitoring security, safety, and other human aspects during requirements specification, design and implementation, and the bearing that this has impact on the generation of trust in software development processes is an area that demands further inquiry.

It was clear from the interviews we performed that there is a critical need for software practitioners to build into their work practises a focus on trust and trustworthiness. Such practices need to be aligned with the differing agendas towards trust and trustworthiness of different stakeholders, including developers, end users, teams, and organisations both building AI-based software and those deploying and using it. There were questions as to whether fundamentally different work practices are needed to achieve the aim of building trustworthiness into AI-based software or whether traditional practices could be used and modified. Similarly, there is much industry and research hype around AI-based software development and issues of fairness, bias, reliability, ethical usage as well as trustworthiness issues. Navigating this is challenging for practitioners building next-generation AI-based software while both technologies rapidly evolve and individual and organisational expectations evolve.

It was apparent from our study findings, aligning with related work discussed earlier, that concepts of trust, trust models, trustworthiness, and definitions, work practices and development tools to support engineering of trustworthy AI-based software are incomplete, or that knowledge of emerging definitions and approaches is patchy. Practitioners indicated that the formation of trust relationships between users and AI, how these relationships are defined, described and evolve, and impact of these concepts on end users of software vary. Participants didn't use theories of trust but tended to base their definitions on their own experiences with technology in their own work and living situations. An interesting finding was developers losing control of how their software might in future be used and not trusting such contexts of use, end users and/or organisations putting it to these uses. Such unexpected or emergent uses of systems, and use of AI-based systems with other, unanticipated systems, may reduce trustworthiness of the AI-based software. The need to assume such unanticipated future uses may or will happen, consider potential future uses, and try and mitigate potential trust impacts is very challenging. A few participants highlighted the need for measures of "trustworthiness" during software development of AI-based systems and the need to address emerging "untrustworthiness" issues. As reported in recent literature, how to define, measure and monitor trustworthiness of AI-based software components is unclear and limited techniques and tools exist. Participants noted that end users could not see into AI-based software on the whole, and that approaches to surface or display concepts that aid in building trust between user and software could be beneficial. What and how to present to users in order to build and maintain trust is unclear for many applications.

Participants and several recent research findings highlight limited treatment of concepts of trust and trustworthiness in software engineering and computer science education and training programmes. Most cover in some detail the requirements, design, engineering, testing and deployment of AI-based software components. Increasingly various ethical concerns relating to such technologies are addressed, including privacy, reliability, explainability and fairness. However, very limited coverage exists of defining trust, identifying individual and organisational concepts of trust, the building of trust relationships, how software developers and development teams build and maintain trust themselves, and how trustworthiness can be achieved.

Conclusion

At the beginning of this article we raised two questions. The first concerned the points made in our discussion in the above section relating to the diversity of definitions and perceptions of trust across academic disciplines, and stakeholders, including as our study has shown software practitioners working in the industry. There is clearly a need to advance a shared understanding of trust which encompasses industry, practitioner and everyday human engagement with AI. Future research might address this question, of how and if it would be possible to generate and disseminate a collective concept and perception of trust across these different layers of theory and practice, and what benefits it might lead to. The purpose of gaining trust is usually embedded in organisational business models, in the technology industry, and as noted above, offered as a service by consultancies. However it is notable that corporate, marketing or UX testing approaches to trust in AI or AI trustworthiness developed either by technology companies or by the organisations were not discussed by our participants; our interviews focused on AI and software applications rather than their views about corporate approaches to trust, however it is also indicative that participants did not raise this issue. We suggest that this layer of tech workers' commitment to trust should be investigated in future research.

Our second question referred to the training and skills that future software practitioners might need to work towards creating trustworthy AI. The answer to this question, we suggest, is related to the first question; we must connect the way that trustworthy AI development is taught to software engineers more closely to the ways that trustworthy AI is experienced, expected and anticipated by those who use it in everyday life. We must also maintain our awareness of the contextual nature of AI applications and the ways that trust and trustworthiness might come about across different contexts. That is, our responses to both of these questions indicate that a new interdisciplinary research agenda for trust in and trustworthy AI is required. We must engage the computing and social sciences together towards a theoretically and practically coherent focus on trust which reframes its treatment as a commodity, towards analysing the possibilities of trust as a connecting thread, between disciplines, stakeholders, practitioners and technologies.

Acknowledgements: Pink is supported by Laureate Fellowship FL230100131. Grundy is supported by Laureate Fellowship FL190100035.

Conflict of interest statement: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data availability statement: Data cannot be made available for reasons of participant confidentiality.

References

Ahmad, K., Abdelrazek, M., Arora, C., Bano, M., & Grundy, J. (2023). Requirements engineering for artificial intelligence systems: A systematic mapping study. Information and Software Technology, 107176.

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

Andras, P. *et al*., (2018) "Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems," in *IEEE Technology and Society Magazine*, vol. 37, no. 4, pp. 76-83, Dec. 2018, doi: 10.1109/MTS.2018.2876107.

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), 6-1.

Belani, H., Vukovic, M., & Car, Ž. (2019, September). Requirements engineering challenges in building AI-based complex systems. In 2019 IEEE 27th International Requirements Engineering Conference Workshops (REW) (pp. 252-255). IEEE.

Bodó, B. & Janssen, H (2021, June 16)., Here Be Dragons – Maintaining Trust in the Technologized Public Sector. Amsterdam Law School Research Paper No. 2021-23, Institute for Information Law Research Paper No. 2021-02, Volume 41, Issue 3, May 2022, Pages 414–429, https://doi.org/10.1093/polsoc/puac019, Available at SSRN: https://ssrn.com/abstract=3868208 or http://dx.doi.org/10.2139/ssrn.3868208

Brusseau, J. (2023) From the ground truth up: doing AI ethics from practice to principles. AI & Soc 38, 1651–1657 (https://doi.org/10.1007/s00146-021-01336-4

Bublin, M., Schefer-Wenzl, S., & Miladinović, I. (2021, September). Educating ai software engineers: Challenges and opportunities. In International Conference on Interactive Collaborative Learning (pp. 241-251). Cham: Springer International Publishing.

Buhnova, B., Halasz, D., Iqbal, D., & Bangui, H. (2023, March). Survey on Trust in Software Engineering for Autonomous Dynamic Ecosystems. In Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing (pp. 1490-1497).

Carleton, A. D., Harper, E., Menzies, T., Xie, T., Eldh, S., & Lyu, M. R. (2020). The AI Effect: Working at the Intersection of AI and SE. IEEE Software, 37(4), 26-35.

Costa, AC, Fulmer, CA, Anderson, NR. Trust in work teams: An integrative review, multilevel model, and future directions. *J Organ Behav*. 2018; 39: 169– 184. https://doi.org/10.1002/job.2213

Desjarlais, R. (2003) Sensory Biographies. California University Press: California.

Ebert, C., & Louridas, P. (2023). Generative AI for Software Practitioners. IEEE Software, 40(4), 30-38.

Estella, A. (2023). Trust in Artificial Intelligence Analysis of the European Commission proposal for a Regulation of Artificial Intelligence. *Indiana Journal of Global Legal Studies*, *30*(1), 39+. https://link.gale.com/apps/doc/A749619739/AONE?u=monash&sid=bookmark-AONE&xid=df0ca13d

Freiman, O. Making sense of the conceptual nonsense 'trustworthy AI'. *AI Ethics* (2022). https://doi.org/10.1007/s43681-022-00241-w

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals,* 14(2), 627-660.

Graetsch, U. M., Khalajzadeh, H., Shahin, M., Hoda, R., & Grundy, J. (2023). Dealing With Data Challenges When Delivering Data-Intensive Software Solutions. IEEE Transactions on Software Engineering.

Hacker, P., Krestel, R., Grundmann, S. *et al.* (2020. Explainable AI under contract and tort law: legal incentives and technical challenges. *Artif Intell Law* **28**, 415–439. https://doi.org/10.1007/s10506-020-09260-6

Hammersley, M & Atkinson, P. (2007). *Ethnography: Principles and Practice*. Taylor and Francis, Hoboken.

Heck, P., & Schouten, G. (2021, May). Lessons learned from educating AI engineers. In 2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN) (pp. 1-4). IEEE.

Hoda, R. (2021). Socio-technical grounded theory for software engineering. IEEE Transactions on Software Engineering, 48(10), 3808-3832.

Howlett, M. (2021). Looking at the 'field' through a Zoom lens: Methodological reflections on conducting online research during a global pandemic. Qualitative Research, 22(3), 387402.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... & Mitchell, M. (2021, March). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 560-575).

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021, March). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 624-635).

Kästner, C., & Kang, E. (2020, June). Teaching software engineering for AI-enabled systems. In Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: Software Engineering Education and Training (pp. 45-48).

Khan, Arif Ali, et al. "AI ethics: an empirical study on the views of practitioners and lawmakers." *IEEE Transactions on Computational Social Systems* (2023).

Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., ... & Akbar, M. A. (2022, June). Ethics of AI: A systematic literature review of principles and challenges. In Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022 (pp. 383-392).

Krüger, S., Wilson, C. The problem with trust: on the discursive commodification of trust in AI. *AI & Society* (2022). https://doi.org/10.1007/s00146-022-01401-6

Laux, J., Wachter, S. and Mittelstadt, B. (2023), Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. Regulation & Governance. https://doi.org/10.1111/rego.12512

Lu, Qinghua, et al. "Software engineering for responsible AI: An empirical study and operationalised patterns." *Proceedings of the 44th International Conference on Software Engineering: Software Engineering in Practice.* 2022.

Lu, Q., Zhu, L., Xu, X., Whittle, J., & Xing, Z. (2022, May). Towards a roadmap on software engineering for responsible AI. In Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI (pp. 101-112).

Martínez-Fernández, S., Bogner, J., Franch, X., Oriol, M., Siebert, J., Trendowicz, A., ... & Wagner, S. (2022). Software engineering for AI-based systems: a survey. ACM Transactions on Software Engineering and Methodology (TOSEM), 31(2), 1-59.

O'Reilly, K. (2011). *Ethnographic Methods* (2nd ed.). Oxford: Routledge.

Ozkaya, I. (2023). Application of Large Language Models to Software Engineering Tasks: Opportunities, Risks, and Implications. IEEE Software, 40(3), 4-8.

Papagni, G., de Pagter, J., Zafari, S. *et al.* Artificial agents' explainability to support trust: considerations on timing and context. *AI & Soc* 38, 947–960 (2023). https://doi.org/10.1007/s00146-022-01462-7

Pant, A., Hoda, R., Spiegler, S. V., Tantithamthavorn, C., & Turhan, B. (2023). Ethics in the Age of AI: An Analysis of AI Practitioners' Awareness and Challenges. *arXiv preprint arXiv:2307.10057*.

Pink, S. (2023) *Emerging Technologies / Life at the Edge of the Future*. Oxford: Routledge.

Pink, S. (2022). Trust, Ethics and Automation: Anticipatory imaginaries in everyday life. In S. Pink, M. Berg, D. Lupton, & M. Ruckenstein (Eds.), *Everyday Automation: Experiencing and Anticipating Emerging Technologies* (pp. 44-58). Routledge. https://doi.org/10.4324/9781003170884-4

Reinhardt, K. Trust and trustworthiness in AI ethics. *AI Ethics* (2022). https://doi.org/10.1007/s43681-022-00200-5

Ritter, C. (2021) Rethinking digital ethnography: A qualitative approach to understanding interfaces. *Qualitative Research*, pp.1-17.

Sajadi, A., *Damevski*, K. & *Chatterjee P. (2023)* Interpersonal trust in OSS: Exploring dimensions of trust in GitHub pull requests. ICSE 2023 NIER - New Ideas and Emerging Results Track. Preprint online at https://preethac.github.io/files/ICSE_NIER_2023.pdf.

Shahrdar, S., Menezes, L., & Nojoumian, M. (2019). A survey on trust in autonomous systems. In Intelligent Computing: *Proceedings of the 2018 Computing Conference*, Volume 2 (pp. 368-386). Springer International Publishing.

Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4), 1-31.

Shostak, M. (2000). *Nisa, the life and words of a !Kung woman.* Cambridge, Mass. :Harvard University Press.

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & Van Moorsel, A. (2020, January). The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 272-283).

Vakkuri, V., Kemell, K. K., Tolvanen, J., Jantunen, M., Halme, E., & Abrahamsson, P. (2022, June). How do software companies deal with artificial intelligence ethics? A gap analysis. In Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022 (pp. 100-109).

Weisz, J. D., Muller, M., Ross, S. I., Martinez, F., Houde, S., Agarwal, M., ... & Richards, J. T. (2022, March). Better together? an evaluation of ai-supported code translation. In 27th International Conference on Intelligent User Interfaces (pp. 369-391).

Widder, D. G., Zhen, D., Dabbish, L., & Herbsleb, J. (2023, June). It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them?. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 467-479).

Williams, R. (2022). Rethinking Administrative Law for Algorithmic Decision Making, *Oxford Journal of Legal Studies*, Volume 42, Issue 2, Pages 468–494, https://doi.org/10.1093/ojls/gqab032.
Zhang, W., Wong, W. & Findlay, M. (2023) Trust and robotics: a multi-staged decision-making approach to robots in community. *AI & Soc*. https://doi.org/10.1007/s00146-023-01705-1

Notes

1. Our focus is on trust and trustworthiness, rather than on the extensive debate on AI ethics, we engage with questions related to ethics here where they overlap explicitly.
2. See for example : https://www2.deloitte.com/be/en/pages/strategy-operations/strategy-analytics-mergers-acquisitions/solutions/trustworthy-ai.html, https://research.ibm.com/topics/trustworthy-ai, and https://ai-watch.ec.europa.eu/topics/trustworthy-ai_en.