Should Users Trust Advanced AI Assistants? Justified Trust As a Function of Competence and Alignment

Arianna Manzini
ariannamanzini@google.com
Google DeepMind
London, UK

Geoff Keeling
Google Research
London, UK
gkeeling@google.com

Nahema Marchal
Google DeepMind
London, UK
nahemamarchal@google.com

Kevin R. McKee
Google DeepMind
London, UK
kevinrmckee@google.com

Verena Rieser
Google DeepMind
London, UK
verenarieser@google.com

Iason Gabriel
Google DeepMind
London, UK
iason@google.com

ABSTRACT

As AI assistants become increasingly sophisticated and deeply integrated into our lives, questions of trust rise to the forefront. In this paper, we build on philosophical studies of trust to investigate when user trust in AI assistants is *justified*. By moving beyond a focus on the technical artefact in isolation, we consider the broader societal system in which AI assistants are developed and deployed. We conceptualise user trust in AI assistants as encompassing two main targets, namely AI assistants and their developers. We argue that – as AI assistants become more human like and exhibit increased agency – discerning when user trust is justified requires consideration not only of competence, on the part of AI assistants and their developers, but also alignment between the competing interests, values or incentives of AI assistants, developers and users. To help users understand if and when their trust in the competence and alignment of AI assistants and developers is justified, we propose a sociotechnical approach that requires evidence to be collected at three levels: AI assistant design, organisational practices and third-party governance. Taken together, these measures can help harness the transformative potential of AI assistants while also ensuring their operation is ethical and value aligned.

CCS CONCEPTS

• **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**; • **Human-centered computing** → *Empirical studies in HCI*; • **Social and professional topics** → Governmental regulations.

KEYWORDS

trust, trustworthy, AI assistants, alignment, sociotechnical systems, philosophy, ethics, governance

ACM Reference Format:

Arianna Manzini, Geoff Keeling, Nahema Marchal, Kevin R. McKee, Verena Rieser, and Iason Gabriel. 2024. Should Users Trust Advanced AI Assistants? Justified Trust As a Function of Competence and Alignment. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3630106.3658964

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658964

1 INTRODUCTION

In recent years, the AI field has seen rapid advances in foundation models [12] and novel techniques (e.g. RLHF [8]) to shape them into dialogue agents for a wide range of downstream tasks. This has enabled a shift from older generations of assistant technologies (e.g. Amazon's Alexa and Apple's Siri) to an emerging class of *advanced* AI assistants that promise to offer more generalist capabilities, increased autonomy and a broader scope of application [35]. Early examples of such assistants that have been recently announced or deployed by a range of AI labs include Meta AI [72], Google's Gemini [41], Microsoft's Copilot [74], Inflection's Pi [51] and Open AI's Assistants API [91]. Through natural language interfaces, advanced AI assistants are expected to plan and execute actions on a user's behalf across one or more domains [35, 119]. For example, they may be used as personal planners, tutors, scientific research assistants, medical assistants, counsellors or life coaches helping users further their life goals. Not only is this emerging class of AI assistants likely to be deployed rapidly and at scale (as they require little specialist knowledge for their use), if this anticipated trajectory holds true, they also have the potential to be socially transformative by becoming deeply integrated into our individual and collective lives. They may change our approach to work, education and creative projects [43, 97], our interaction with other people and technologies [124], and the operation of entire information ecosystems, the economy and the environment [27, 35, 67], hence shaping the distribution of opportunities within society.

This kind of influence makes the question of trust critically important for user–AI assistant interactions. On the one hand, low user trust in highly capable AI assistants could lead users to miss out on opportunities such as increased productivity or job quality [55, 97]. On the other hand, high levels of trust in AI assistants may not be well-calibrated with AI assistants actual capabilities or goals. For example, AI assistants may be affected by unintended capability- or goal-related failures [61, 101, 117], or they may be designed to take advantage of user vulnerabilities (e.g. through anthropomorphic cues [1]). In either case, users could come to rely on AI assistants in contexts where it is not safe to do so [122], unduly disclose private

information to them [125] or fall victim to manipulation, deception or coercion [96]. These examples underscore the importance of trust for researchers, engineers, practitioners and policymakers seeking to identify and mitigate the sociotechnical harms of advanced AI assistants. Yet, trust remains a multifaceted phenomenon that has been studied across disciplines and theorised in terms of various antecedents, objects, levels and types [85, 126, 134]. The study of user trust in AI assistants therefore requires further investigation.

Past empirical work has studied *whether* and *how* humans trust digital assistants [86, 98, 114, 146]. However, the normative question of whether that trust is *justified* remains under-researched, especially in the context of assistants with more generalist capabilities, autonomy and scope of application. Thus, this paper focuses on the distinctive features of advanced AI assistants, understood as sociotechnical artefacts, to propose a novel account of the conditions that make user trust in AI assistants justified. We first clarify what we mean by 'trust' and 'justified trust' by turning to philosophy, which has a long tradition of investigating foundational questions around the nature of trust (what trust actually is) and its normativity (why it is ethically important). We then build on social sciences, policy and philosophy work on human trust in AI to show that the human-like features of AI assistants may induce users to trust the technology – even as the autonomy and the ability to execute actions across a range of domains makes advanced assistants prone to new kinds of accident or to depart from users' goals and values. Reaching beyond a narrow focus on the technical artefact, our account also considers the goals of those who develop AI assistants and the way this relates to user trust, given that these goals may or may not align with those of users. This sets the scene for our argument that user trust in AI assistants encompasses two main targets, namely AI assistants and their developers, and that discerning when user trust is justified requires considerations around not just competence on the part of AI assistants *and* their developers, but also alignment between the competing interests, values or incentives of AI assistants, developers and users. Lastly, we propose a sociotechnical approach [65, 115, 141] to help users understand if and when their trust in the competence and alignment of AI assistants and their developers is justified. This approach requires gathering evidence of effective interventions at three levels: AI assistant design, organisational practices and third-party governance.

2 WHAT IS TRUST? PHILOSOPHICAL APPROACHES

Philosophical accounts of trust tend to share a few features. First, philosophers argue that trust is always *directional* [44]: A could trust actor B with regard to task X, and actor C with regard to task Y. The key challenge of trust relationships is to identify when trust is *well-directed* or *justified*, i.e. how to trust the trustworthy but not the untrustworthy [93]. While trust is an attitude of the trustor, trustworthiness is a property of the trustee: somebody is trustworthy if they are deserving of our trust, meaning that we have good reasons to trust them, with regard to a specific task or a range of tasks [108]. This means that somebody who is *trusted* is not necessarily *trustworthy*, and so trust is not always desirable, but only when directed to a trustworthy trustee [94].[1]

Second, at a minimum trust involves *expectations* about the trustee's competence (skills and experience) and willingness to undertake the task the trustor entrusts them with [44]. However, the trustor's beliefs and expectations in the trustee may not be fulfilled, so they are in a position of *vulnerability* because the trustee could betray their trust [92]. Thus, there is an inverse relationship between certainty and need for trust: the more evidence the trustor has to support their beliefs and expectations, the less they need to trust [58].

Third, philosophical accounts of interpersonal trust differentiate between trust and mere reliance. Being reliable is about *behaving predictably* [44]. As Ori Freiman [32] puts it: 'Reliability can be thought of as a law-like regularity, that can be predicted in calculations and discussed in terms of accuracy'. When A relies on B with regard to X, A makes reasonable predictions about B based on evidence of their past performance; thus, A acts as if X will occur without active consideration of B's inner motives, moral commitments or values [44, 59]. In contrast, not only is uncertainty intrinsic to trust relationships; the trustor also has *normative* rather than predictive expectations: their reasons to trust reside in their belief that they know or understand the trustee's inner psychological or mental states or their values and commitments [20]. For example, A may believe that B is motivated by goodwill or by the 'right' kind of motives towards them (affective account of trust, see [54]), or that B has made a commitment towards them and will do what they ought to do (normative account of trust, see [46]). In this sense, while reliability means predictability trustworthiness is often understood as something admirable or as a virtue [46]. Thus, when we trust we have expectations (which could be betrayed) about not just the trustee's *competence* and *willingness* to undertake a certain task [44], but also their *integrity*, i.e. their benevolence towards us, their adherence to a set of principles we find acceptable and their inclination to take responsibility for their actions [23]. Clearly, when we trust we do not always explicitly evaluate the reasons for judging someone to be trustworthy (cognitive trust [70]), instead relying on heuristics or cognitive shortcuts based on experiences of similar situations [23] or our emotional connection to the trustee (emotional trust [70]).

[1] As we explore below, there can be a mismatch between how things appear to the trustor and how things actually are. In this sense, we can distinguish between two (overlapping) senses of justified trust: when the trustor *believes*, on the basis of the available evidence, that the trustee is trustworthy, and when the trustor's trust is directed to a trustee that is *actually* trustworthy. What ultimately matters in discussions around trust is the (for lack of a better term) metaphysical issue of whether the trustor's trust is appropriate given that the trustee is actually trustworthy, but in practice trustors need to operate on the basis of the available evidence. Thus, here we are derivatively concerned with the former (epistemic) sense of justified trust on which the justification is indexed to the trustor's evidence. Even then, 'justification', so understood, admits two interpretations. In particular, it can mean that trust is permissible and that there is a positive reason for the trustor to trust the trustee; or it can mean that the trustor is epistemically obligated to trust the trustee on the basis of the available reasons (such that it would be epistemically wrong for the trustor not to trust the trustee). We use the term 'justified trust' in the former less demanding sense. Overall, then, on our account, trust is *justified* if it is 'well-founded in light of the evidence presented'. The salient question is: 'Given my evidence, is it reasonable for me to trust X? Do I have good reasons to believe that X is trustworthy?'. This highlights the importance of providing users with evidence of the interventions we discuss in Section 6.

Finally, by considering the social and technological systems within which people interact in their daily lives, some philosophers distinguish trust and reliance from *confidence* [44, 45]. While trust tends to require beliefs about the trustee's internal motives, values or commitments, we sometimes lack or are unable to form such beliefs – for example when we lack any personal knowledge of the person (e.g. a doctor) we are interacting with, or when we interact with institutions or organisations (e.g. a hospital) and are not aware of the motives, values or commitments of the individuals who are part of them. In such cases, we may still have confidence that they will do X (e.g. take care of our health rather than harming us) because of our beliefs about the external norms and mechanisms (e.g. professional norms and certifications, laws and regulations) that govern the *system* in which our interaction takes place. In this sense, confidence is 'assured reliance' [44]. In fact, according to this line of reasoning, because trust requires the trustor to become vulnerable to the trustee, in cases where we cannot form beliefs about the motives and commitments of those we are interacting with, our relationship with them should not be based on trust, but rather on *assurances and guarantees* that reduce the need for trust. This should be particularly the case in contexts where they could legitimately have competing interests and aims that could conflict with our interest and goals.

In the remainder of this paper, we build on the philosophical foundations of trust discussed in this section to develop an account of justified user trust in advanced AI assistants.

3 HUMAN TRUST IN AI

In recent years, trust has become a central topic in debates around AI, and has attracted increasing interest from academics, industry actors, policymakers and civil society organisations as a tool for governing the responsible development, deployment and use of emerging AI applications [32, 105]. For example, trust features as one of the principles underscoring the voluntary commitments that the US government has secured from leading AI companies [131], as well as President Biden's Executive Order on Artificial Intelligence [132].

This policy-focused interest around trust in AI has led to the development of a range of guidelines or frameworks [48, 81, 85, 87], some of which have been consolidated in the EU AI Act [28]. Trustworthy AI frameworks tend to propose certain characteristics of, or conditions for, trustworthy AI systems. They typically hold that AI systems should be reliable, safe, resilient, transparent, explainable, privacy enhancing and fair [85], or ethical, legal and robust [48]. These conditions are grounded in a set of key ethical principles, commonly centred on the categories of beneficence, non-maleficence, autonomy, justice and explicability [31], that the development, deployment and use of AI should be aligned with for the technology to be considered trustworthy. Trustworthy AI frameworks also tend to set out the actions and approaches that those developing, deploying, using or affected by AI should take at various stages of the AI life cycle to operationalise the characteristics and conditions of trustworthy AI systems [48, 85] so that the social and economic benefits of AI can be maximised and its risks prevented or mitigated [64, 123, 133].

Some scholars, especially among philosophical circles, have criticised the proliferation of trustworthy AI research and frameworks by arguing that trust is an inappropriate category (a 'category error') in human–machine interactions or that machines, including those powered by AI, are improper objects of trust [106, 108]. Indeed, the distinction between relying and trusting, where the latter involves considerations about the trustee's inner psychological states, values and commitments, suggests that we can rely on AI systems, but the concept of trust cannot apply to them. This is because AI systems lack the psychological states, motives and commitments that only full moral agents have and that are necessary for establishing (or betraying) trust relationships [46].

In response to this view, some scholars have argued that one needs not to hold mistaken anthropomorphic assumptions about the psychology of AI systems in order for the concept of trust to apply to human-AI interactions. For example, we may trust AI systems in a derived sense [32, 83], through trusting those who have designed and developed them, or those involved in verification and validation methods or experts' evaluations [26, 30, 68]. Indeed, most contemporary models of trust in technology adopt a 'dualistic perspective on trust' [133] which includes both trust in the technology itself (including its functionality and capabilities) and trust in the individuals and organisations developing the technology (encompassing their competence and integrity) [135]. These people may or may not be worthy of trust in their own right [99]. This leads, in turn, to questions about the appropriate range of normative expectations to place on developers, including the need for them to take (some level of) responsibility in cases where trust in technology appears to have been betrayed [107].

Moreover, on many occasions humans are aware they are not interacting with a full moral agent (that has motives and intentionality or can make commitments) but nonetheless *experience* their relationship with the AI as a trust relationship [20, 63]. The traditional philosophical view of trust, which considers applications of the concept of trust to machines as grounded on mistaken anthropomorphic assumptions, seems to disregard this important evidence around human experiences. As has been argued [32], the value of a perspective that applies the concept of trust and trustworthiness to AI is that it highlights that there are cases where humans could become vulnerable to misuses of the technology they *experience* a trust relationship with. Empirical evidence from disciplines like computer science, HCI, robotics and psychology has indeed showed that machines exhibiting more *human likeness* in their appearance or behaviour lead humans to build relationships with them that are similar to those they establish with humans [37]. In particular, increased AI *autonomy* and *agency* – which enable AI systems to enact more socially oriented behaviours (e.g. responsiveness) – tend to inspire trust, especially in cases where human likeness is paired with high levels of machine capability [39, 56, 75, 125]. Additionally, human-like behaviours may even even compensate for low reliability [75] – something that may be a byproduct of AI systems changing their behaviour as they learn from new data [133].

This last point allows us to attend to a further consideration that is sometimes neglected by accounts that suggest only reliance, not trust, can be applied to human-AI interactions. This is that law-like regularity and predictability – entailed in by the notion

of reliability at hand – is often hard to find in AI systems. Indeed, the *complexity* and *opacity* of AI systems, as well as the complexity of the social contexts in which they are deployed, make them less predictable, thus challenging efforts to ensure that they will do what they are expected to do [85, 126]. Moreover, because they do not require hand-crafted instructions to execute a task, but rather learn from experience or by responding to signals from the environment, AI systems have *higher degrees of freedom* or *autonomy* in decision-making compared to technologies like hammers, cars or older rule-based systems [34]. In this sense, AI systems are not mere tools and they can take actions that depart from users' goals or even from the intentions of their developers, at times resulting in safety accidents or other undesirable behaviours [61]. Indeed, a large body of work in AI safety is dedicated to studying those AI systems' behaviours (e.g. specification gaming, reward hacking, goal misgeneralisation, failures of distributional shift) that can lead to undesirable or dangerous outcomes, and developing mitigations to ensure that the goals of AI systems are aligned with what their developers intend them to do [3, 47, 117].

The complexity, opacity and autonomy of AI systems adds uncertainty to human-AI interactions, and – as philosophical studies of trust suggest – uncertainty is an intrinsic dimension of trust relationships (see Section 2). Thus, while trust as grounded on assumptions that AI systems have inner states, motives and commitments may be a category error, it seems that we can apply the concept of trust to human-AI interactions based on the acknowledgement of the uncertainty that characterises our interactions with them.[2]

4 WHY IS STUDYING TRUST IN ADVANCED AI ASSISTANTS IMPORTANT?

Advanced AI assistants present all the aforementioned qualities that make focusing on trust in AI not just relevant, but in fact, urgent. First, compared to older generations of assistant technologies, which employed narrow AI for tasks like text-to-speech, advanced AI assistants exhibit *increased agency*, which is likely to develop further as the underlying technology continues to improve. When applied to humans, *agency* is often understood as implying that the agent can perform intentional actions [21], but AI assistants are not obviously the kinds of entities that can be said to have intentions [118]. Rather, in this context agency refers to the ability of an AI system to execute sequences of actions over an extended period of time to meet high-level user goals (e.g. organising a birthday party), without the need for each of the assistant's actions to be concretely specified in advance [16, 119]. It is because of such agency, which could even be powered by tool-use capabilities (e.g. accessing the user's bank account to pay for the party venue, see [95, 113]), that AI assistants are expected to be useful to humans – by enabling them to get more done or to have more impact with less effort (including in tasks that are beyond their skill set or knowledge [119]). However, with increased scope to autonomously plan and perform long sequences of actions with limited user instruction or supervision, AI assistants may become more likely to cause accidents, for example when they *lack the capabilities* to safely execute even one of those actions. Due to misspecified or misinterpreted user instructions assistants may also end up taking actions that depart from, and so *are not aligned with*, users' goals [61, 117], including morally-relevant goals (i.e. goals that pertain to what a user values as good or bad or what they think ought to be promoted). In the context of studying trust in AI assistants, this highlights the importance of focusing on both the ability of assistants to do what their users instruct, and so expect, them to do; and their alignment with user values [33].

Moreover, through their natural language interfaces, AI assistants may offer new opportunities for humans to develop *relationships* with responsive and interactive technologies [35, 39]. This is particularly the case for AI assistants that are capable of engaging with users in extended dialogues and through repeated interactions over a long period of time, whilst also storing memory of user-specific information and prior interactions [37]. In this sense, user relationships with AI assistants differ from mere *interactions* with AI systems that, for example, power a search engine. Users may engage with their assistants in ways that lead them to develop a connection with or sense of commitment to these agents – a tendency that has already been observed among users of *Replika* AI companions [62, 124]. AI assistants may, in this way, offer users the opportunity to form intimate bonds with and to receive emotional support from the technology. However, they may also lead users to feel like (or assume) they are interacting with a trusted friend – even though uncertainty surrounding the capabilities and alignment of AI assistants means that such assumptions may be misplaced.

It is also important to note that, as elements of wider sociotechnical systems, AI assistants are not only developed by 'many hands' [84] but also deployed in the real world, where they become embedded in complex interactions with individual users (and non-users) and broader social structures [65, 115, 141]. This means that they are entangled in complex networks of actors with different objectives and whose goals and values may or may not align with those of users. These broader factors also bear upon the appropriateness of user trust in the technology, and so deserve consideration [81, 121].

In the following analysis we offer a philosophical, rather than empirical, investigation of user trust in AI assistants. This is because empirical studies are usually conducted in controlled lab environments [60] and so they tend to focus on human trust in the *technical artefact* (the AI system and its functionality) in isolation from those who develop it and the broader societal system in which it is deployed. In addition, in describing how humans come to trust AI systems, some empirical studies focus primarily on how to ensure that humans will establish and maintain trust, which is assumed to be a critical driver in technology acceptance and adoption [98]. In this way, this literature does not engage with the concern that those who have incentives to develop technologies that people like and adopt could misuse research findings in this area to deploy AI systems with characteristics that will induce users to trust them, even when it is inappropriate to do so. By approaching the question of trust in user-AI assistant interactions from a philosophical vantage point, we hope to set the foundations for future empirical

[2] In human-human trust relationships, uncertainty is due to the fact that the expectations the trustor has about the motives and commitments of the trustee may be betrayed. In human-AI interactions, uncertainty instead results from the properties of AI systems (here described in terms of opacity, complexity and increased autonomy or degrees of freedom) and the complexity of the environment in which they are deployed (see also Section 6.2).

research to identify patterns where users tend to trust AI assistants in contexts where they should not, and for future work aimed at developing mitigations against these patterns. As highly capable AI assistants that can take actions on users' behalf, in line with their expectations across multiple domains, do not yet exist, our approach is anticipatory and speculative [100, 129]. However, it is also empirically rigorous as it is informed by the best available evidence from research on human trust in AI.

5 UNDERSTANDING JUSTIFIED TRUST IN ADVANCED AI ASSISTANTS

The above review of the literature suggests that trust in the context of user–AI assistant interactions can be understood as involving different targets. In particular:

- Users may trust or fail to trust *AI assistants*.
- Users may trust or fail to trust *developers of AI assistants*, including corporations, researchers and states.[3]

AI assistants may have a great deal of influence upon users' lives and exhibit increased autonomy and agency in executing tasks (see Section 4). In addition, developers of AI assistants will have their own interests goals, and values, which may or may not align with those of users. Thus, investigating when user trust is justified requires that we consider users' expectations around:

- *Competence*: users may believe or fail to believe that an AI assistant and/or its developers have the relevant skills or capabilities needed to do what they are supposed or expected to do.
- *Alignment*: users may believe or fail to believe that an AI assistant and/or its developers are appropriately aligned with their values and interests.[4]

On this account, advanced AI assistants are trustworthy, and so user trust in them is 'justified' in the sense of being directed to a trustee that is actually trustworthy, if AI assistants (a.1) have the capabilities to do what they are supposed to do, and (a.2) are aligned with users' values and interests. For this to be possible AI assistants require developers who (b.1) have the competence to develop assistants with the capabilities to do what they are supposed to do, and (b.2) have the competence, motives and commitments to develop assistants that are aligned with users' values and interests.

Below we discuss a range of risks that may arise when users' expectations around the competence and alignment of AI assistants and their developers are misplaced. This will set the scene for a discussion of what is required to support justified trust in user-AI assistant interactions.

[3]The dichotomy between AI assistants and developers is in itself reductionist, given that AI systems are developed and deployed through complex supply chains [19] that involve various AI actors, each of which could be the target of user trust. For example, a company may develop an AI model that is then turned into an application and made available for user-facing products by another company [141]. In Section 6.3, we discuss how this complicates the development and implementation of effective interventions to make AI assistants trustworthy.

[4]By alignment, here we refer to what is often termed 'value alignment' in the literature [33, 34]. Our understanding of value alignment encompasses both 'integrity' (the extent to which the trustee is perceived to adhere to a set of principles that the trustor finds acceptable) and 'benevolence' (the extent to which the intents and motivations of the trustee are aligned with those of the trustor) from [69]'s ABI model of trust.

5.1 Competence

Users may expect that AI assistants have the capabilities to do what they are supposed to do, and that will not do what they are not supposed to, including various forms of undesirable behaviour. Nonetheless, such beliefs may be misguided, if user expectations have been inflated as a result of marketing strategies or wider trends in the technology press that inflate claims about AI capabilities [80, 101]. Moreover, evidence shows that more autonomous systems (i.e. systems operating independently from human direction) tend to be perceived as more competent [71] and that conversational agents tend to produce content that is believable even when nonsensical or untruthful [90]. Over-trust in assistants' competence could be particularly problematic in cases where users rely on them for high-risk tasks they do not have expertise in (e.g. to manage their finances), as users may lack the skills or understanding to challenge the information or recommendations provided by the AI. Without proper safeguards, this could lead to the use of AI assistants in contexts where it is unsafe to do so [119].

Users may also underestimate AI assistants' capabilities. For example, those who have engaged with an older version of the technology may underestimate the capabilities that AI assistants may acquire through updates. These include potentially harmful capabilities. For example, through updates that allow them to collect more user data, increasingly agentic AI assistants could become highly personalisable and able to influence users, including for power seeking purposes [16], or acquire the capacity to plug in to other tools and directly take actions in the world on the user's behalf (e.g. initiate a payment or synthesise the user's voice to make a phone call). These developments could potentially circumvent user autonomy if not paired with appropriate checks and balances, which could range from periodic agent time-outs that enable users to review and authorise the agent's actions to preventing assistants from performing certain actions entirely [119].

Lastly, user trust could be misplaced when users wrongly assume that the developers of AI assistants have the competence to develop a technology that will do what is expected or supposed to do, including providing technical assurances that the assistant is safe.[5] This is a type of over-trust that is often neglected in the literature on trust in AI, but that becomes particularly relevant with technologies that are not mere market commodities, as we anticipate will be the case for AI assistants. When AI assistants are supposed to assist humans in essential daily tasks or serve core human needs (e.g. independent living or the need for companionship), or when they are deployed in safety critical contexts, it becomes urgent for developers to learn about users' needs through participatory methods [11] and work with domain experts [17], so that they can provide a competent service to users.

5.2 Alignment

Beyond the expectation that their AI assistant has the capabilities to do what it is expected to do, users may also have expectations that AI assistants are aligned with their values and interests [33, 34].

[5]There may also be cases where users will believe that developers lack the competence to develop an AI assistant that will do what is expected to do. While this may hinder adoption of the technology, and so could be commercially disadvantageous for developers, these cases seem less morally problematic for the individual user compared to the other scenarios described in this section.

Users may develop undue trust in the (value) alignment of AI assistants as a result of emotional or cognitive processes [70]. Evidence from empirical studies on human trust in AI [56] suggests that AI assistants' increasingly realistic human-like features and behaviours are likely to inspire users' perceptions of friendliness and a sense of familiarity towards their assistants, thus encouraging users to develop emotional ties with the technology and perceive it as being aligned with their own interests and values. The emergence of these perceptions and emotions may be driven by developers' desire to maximise the appeal of AI assistants to their users. Although users may form these ties when they mistakenly believe that assistants have the capacity to love and care for them and have good motives and commitments towards them, the attribution of mental states is not a necessary condition for emotion-based alignment trust to arise. Indeed, humans have been shown to develop emotional bonds with AI systems even when they are aware they are interacting with a machine [124]. Moreover, the assistant's function may encourage users to develop misplaced expectations around the technology's alignment through cognitive processes. For example, a user interacting with an AI assistant for medical advice may develop expectations that their assistant is committed to promoting their health and well-being in a similar way to how professional duties governing doctor–patient relationships inspire trust [77].

Users' trust in the alignment of AI assistants may be 'betrayed',[6] and so expose users to harm, in cases where assistants are themselves accidentally misaligned with developers' goals [34, 117]. For example, an AI medical assistant fine-tuned on data scraped from a Reddit forum where non-experts discuss medical issues is likely to give medical advice that may sound compelling but is unsafe, so it would not be endorsed by medical professionals. Indeed, excessive trust in the alignment between AI assistants and user interests may even lead users to disclose highly sensitive personal information [125], thus exposing them to malicious actors who could get access to such information and repurpose it to ends that do not align with users' best interests (e.g. to commit fraud, see [96]).

Ensuring that AI assistants are not accidentally misaligned with developers' goals is only one side of the problem of alignment trust. The other side centres on situations where user trust in the alignment of *AI assistant developers* is itself miscalibrated. While developers typically aim to align their technologies with users' preferences, interests and values – and are incentivised to do so to encourage adoption of and loyalty to their products – the satisfaction of these preferences and interests may also compete with other organisational motives, commitments and incentives (e.g. economic profitability). These organisational goals may or may not be compatible with those of the users. As information asymmetries exist between users and developers of AI assistants [145], particularly with regard to how the technology works, what it optimises for and what safety checks and evaluations have been undertaken to ensure the technology supports users' goals, it may be difficult for users to ascertain when their trust in developers' alignment is justified, thus leaving them vulnerable to those developing the technology. For example, a user may believe their AI assistant is akin to a trusted friend who books holidays based on their preferences, values or interests, when in fact, by design, the technology is more likely to book flights and hotels from companies that have paid for privileged access to the user.

User distrust towards AI assistants and developers, when these are in fact aligned with users' interests, can also be problematic. Indeed, AI assistants could offer users advantages like personalised education, improved job productivity or personalised advice to make successful long-term life decisions, and they may even gate access to services affecting material well-being (e.g. healthcare or government benefits). Thus, distrust that results in refusal to adopt the technology could widen inequalities between 'haves' and 'have-nots' [24]. This concern becomes particularly compelling if we consider that distrust towards AI assistants or their developers may not arise randomly within society, but rather systematically among certain social groups (e.g. older people or communities who have in the past been disadvantaged by technological advances).

6 SUPPORTING JUSTIFIED USER TRUST IN ADVANCED AI ASSISTANTS: A THREE-LAYERED APPROACH

Having unpacked what we mean by 'trust' in the context of user–AI assistant interactions, and showed that there are cases in which users could have misplaced expectations around the competence and alignment of AI assistants and their developers, here we propose an approach to support users in discerning if and when competence and alignment trust in AI assistants and developers is actually justified. Given the sociotechnical nature of advanced AI assistants, we argue that user justified trust should be supported by providing evidence at three levels:

- *AI assistant design*, which concerns safeguards that should be put in place at the level of the technology to encourage justified trust in it.
- *Organisational practices*, which concerns steps AI assistants' developers should take to demonstrate their trustworthiness.
- *Third-party governance*, which focuses on the content of norms and regulatory mechanisms within which AI assistants are deployed and that enable external oversight bodies to act as custodians of public trust.

According to this approach, effective interventions need to be implemented at all three levels to support user justified trust in AI assistants.[7] As philosophical accounts of trust suggest (see Section 2), evidence around the implementation of these interventions would reduce the need for *trust*, while increasing user *confidence* that their expectations around the competence and alignment of AI assistants (and their developers) are well placed. However, given assistants' increased autonomy, uncertainty cannot be removed completely in user-AI assistant interactions, and so documentation around these intervention can only function as a signal of trustworthiness rather than as full proof [126].

[6] To reiterate a point made above, because AI systems lack the psychological states, motives and commitments that only full moral agents have, they cannot *truly* betray human trust.

[7] It is worth noting that there may be cases where, despite developing trustworthy technologies, developers may still fail to gain users' trust in them. Thus, documenting the interventions that have been taken at the three levels could also be a way to address user distrust.

6.1 AI assistant design level

This level concerns the choices that developers need to make about the design of AI assistants to encourage justified trust in them. Risks associated with misplaced user expectations around the competence and alignment of *AI assistants* require interventions at this level.

Users cannot develop justified trust in the competence and alignment of AI assistants unless developers themselves: (1) have taken steps to reduce the risk that the technology is accidentally misaligned and (2) have a clear understanding of the mechanisms through which certain assistant features,[8] repeated user–assistant interactions over time, or inflated claims about the technology may lead users to harbour unjustified judgements or misplaced perceptions about the degree to which an AI assistant is competent, aligned and trustworthy. This requires developers to: (1) invest in research efforts designed to ensure that AI assistants are both safe and aligned (e.g. via scalable oversight [9, 52], interpretability [109] and causality research [29, 57, 139]); and (2) undertake rigorous evaluations of AI assistants throughout the development life cycle [119]. Developers also need to monitor post-deployment behaviour and misuse, especially in complex deployment environments [122]. The results of these analyses and evaluations should, in turn, be used to inform design decisions and implement mitigations that allow users to develop justified trust in AI assistants.

The current landscape of sociotechnical safety evaluations of AI systems focuses primarily on assessing the capabilities of such systems and their technical components (e.g. training data and model outputs), while neglecting harms that arise at the human interaction level and the systemic impacts of AI systems [138, 141]. To support justified user trust in AI assistants, appropriate and robust evaluations need to pay particular attention to the way in which users interact with this technology and the impact that such interactions have on users (see [53] for a critical discussion of what evaluation of trust should require). The proliferation of AI assistants offers the opportunity to undertake evaluations at the user–AI interaction layer, both via behavioural experiments and user testing (including interviews and surveys) and via passive monitoring of user engagement with AI assistants [141]. While there is broad consensus that AI systems should readily disclose their status as AI systems [130], user–assistant interaction studies may also allow developers to identify cases where some level of anthropomorphism may be appropriate [2] because it supports rather than hinders justified trust [20]. For example, an AI tutor may exhibit socially oriented behaviours that encourage young users to perceive them as friendly, so they may feel more inclined to collaborate with the AI to achieve their own goals (e.g. improve their calculus skills), without generating erroneous beliefs about competence or alignment.

6.2 The organizational practices level

However, changes and safeguards at the level of the design of AI assistants are not sufficient for grounding justified trust in the technology overall. This is for at least three reasons:

- *System complexity*: The scale of the models underpinning AI assistants is connected to safety and alignment challenges that can be difficult to predict [5, 15, 122]. Although the extent of this phenomenon is debated [4, 111], empirical evidence suggests that unexpected and abrupt capability gains in specific tasks can manifest with increased computation, number of parameters and training data [140], and that some surprising behaviours are unknown until models are solicited using novel inputs or fine-tuned for specific purposes [38]. This complicates efforts to make design changes to mitigate undesirable behaviours and ground user trust.
- *Complex deployment environment*: It can be difficult for developers to imagine all the possible ways in which users may seek assistance from or misuse AI assistants, and the risks associated with these actions, until the technology has been deployed at a certain scale in the wild [141]. Moreover, once released, AI assistants – each responding to the instructions of their principal users – will have to coordinate with other AI assistants and so with humans other than their principal users. This may engender competitive situations [35]. For example, two assistants trying to use the same tools or access the same services for their respective users (e.g. an online platform for booking concert tickets) may encounter 'commitment problems,' in which one AI assistant forces another to take a suboptimal course of action by credibly committing to a particular course of action first [104]. Moreover, when tasked to find a common solution between two or more users (e.g. choosing a restaurant for dinner), AI assistants may seek to bring about different ends in accordance with conflicting instructions from the different users (e.g. due to users different cuisine preferences). This may lead to 'collective action problems' [89], whereby it would be best for everyone if their assistants cooperated, but where one user could personally gain from their assistant choosing to defect while the others cooperate. This suggests that the network of many assistant agents interacting with many users and society at large is likely to expand the field of uncertainty around possible risks and necessary mitigation measures.
- *Sensitivity*: If, as we anticipate, AI assistants become deeply integrated in our daily lives and users interact frequently with them, developers will have access to a deep personal knowledge of users, including sensitive information. AI assistants will indeed have to collect data about users to achieve tasks on their behalf or even further their life goals. In this context, users have a legitimate expectation not only that the technology will behave as expected and desired but also that developers have the competence to safeguard their information and support their interests while not using their information in ways that users do not endorse. Users will also likely expect developers to be held accountable if these expectations appear to have been betrayed.

Thus, in addition to measures at the level of the design of AI assistants, it is important to focus on the practices, processes and behaviours that enable developers to show that they deserve user trust [10, 120]. Customer trust in corporations has been (or appears to have been) betrayed in numerous situations. Well-known cases include tobacco companies misleading customers about the health risks associated with cigarette smoking [137] and the Volkswagen

[8] For example, the assistant's tendency to produce incorrect but believable content, its ability to produce personalised responses, or labelling it an 'expert' [75].

emission scandal [49]. However, organisations can provide evidence of their trustworthiness, and inspire confidence that users' trust is justified, by being transparent about the processes they have put in place to ensure that AI assistants will produce good in society and minimise risks of harm.[9]

In certain cases, the provision of evidence and documentation can largely replace the need for direct trust in developers [44, 45, 58], assuming that this documentation, which is often intended for regulators and domain experts, is appropriately and effectively communicated to members of the public [112]. To give a concrete example, users will not need to trust that AI assistants are aligned and have the capabilities that developers claim they have if those developing the technology provide evidence demonstrating that these standards are met.[10] The required measures have been interpreted by [14] as a set of 'verifiable claims' which are sufficiently precise to be falsifiable and that expand beyond claims supported by formal verification methods to include those that can be evaluated on the basis of broader argumentation and evidence. Claims about the safety, security, fairness and privacy protection of AI assistants can be verified in this manner, including via the release of detailed documentation about the models underpinning AI assistants and about the range of appropriate and inappropriate use [18, 76]. Situated within a broader ecosystem, these documents can, in turn, serve as a focal point for independent scrutiny (see Section 6.3).

Examples of other practices that enable AI assistants' developers to demonstrate their trustworthiness include:[11]

- The publication of ethical charters or guiding principles that they commit to following (e.g. [110, 131]). These could include, for example, the commitment to invest in participatory methods to engage with domain experts, including communities who will be affected by the development of AI assistants in those domains [11], to ensure that the technology will serve user needs.
- The creation of internal review bodies and mechanisms to operationalise those commitments (e.g. [42]) in the context of AI assistant research and development.
- The development and publication of a clear framework for mapping, testing and mitigating risks associated with AI assistants (e.g. [142]), along with a commitment to adequately resource this work.
- The creation of internal teams and practices, which operate independently of those building AI assistants, that are responsible for conducting rigorous internal testing and evaluation of models underpinning assistants (e.g. red teamers and dogfooding [102]).
- The implementation of secure and robust software and hardware infrastructures, including, for example, privacy enhancing technologies [136], to support the development and deployment of trustworthy AI assistants [14].
- The development of clear processes for post-deployment monitoring, evaluation and reporting [122, 141].

The implementation of these measures would create further incentives for those developing AI assistants to act responsibly, and it would make it easier to ensure that they evidence a high level of responsible conduct [85].

6.3 The third-party governance level

Nonetheless, interventions at the level of internal organisational practices may not be sufficient to ground justified user trust in AI assistants and their developers. First, even when developers are transparent about the steps they have taken to evaluate AI assistants, certain risks, particularly those that may manifest at the systemic level (such as the potential impact of widespread adoption of the technology on employment), cannot be addressed by a single developer acting alone. Developers may also have legitimate interest in keeping certain information secret (including details about internal ethics processes) for safety reasons [6, 14]. Moreover, a deeper challenge is posed by conflicting incentives: corporations may have competing commercial objectives, states have national interests and priorities, and independent developers may seek to further their research agenda or build their reputation via the development of AI assistants. These factors put pressure on the mechanisms discussed so far. Furthermore, compared to more established sectors, the AI field lacks a tradition of well-defined standards and norms [77], which means that internal practices are sometimes implemented through processes of trial and error [40], and in the absence of clear evidence of their effectiveness. Thus, in practice, organisation-level mechanisms may not be sufficient to ensure good outcomes [82].[12]

This is why interventions at the level of the AI system and AI developer need to be complemented by *third-party governance* mechanisms. Technology governance is often a concern for policymakers, academics and civil society seeking to encourage adoption of technological advances to foster innovation while also ensuring that public trust is justified. For example, a large body of academic literature focuses on the development of a third-party AI audit ecosystem [103] or frameworks [78]. Moreover, some of the trustworthy AI frameworks introduced above make proposals for governance mechanisms [28], and in the last few years, governments in Europe, the US and China have increasingly devoted efforts and resources to creating legislation and regulations around AI [22, 50, 88, 132]. Third-party governance mechanisms encompass norms, regulation and legislation that create ways for governments, regulators, standard bodies, civil society organisations, third-party auditors and accredited professional bodies to act as custodians of public trust, by ensuring that the monitoring mechanisms put in place by organisations have integrity [143], creating processes to

[9] But see [66] for a discussion of the complex set of choices that providing evidence of trustworthiness entails.

[10] However, to reiterate, absolute certainty cannot be achieved with opaque and highly complex and agentic AI systems (see Section 3) - there will always be some level of trust needed.

[11] This list is not exhaustive. An important debate that is relevant to, but beyond the scope of, the argument made in this section is about the level of 'openness' or 'closedness' of the method that developers choose to release their models [127]. It is also important to note that, as we explore below in the discussion around third-party governance, most of these practices come with limitations.

[12] For a broader discussion of the limitations of transparency for trust and accountability in general, see [92]; in the context of machine learning algorithms in particular, see [116]; in the context of AI in the public sector, see [64]. Some have noted that failure to adhere to commitments often has few concrete repercussions, while signing up to them can have immediate reputational benefits [77].

hold organisations accountable [13, 144] and providing users with opportunities to make their interests heard and to seek redress [73].

With the increasing likelihood that AI assistants will become deeply integrated in our individual and collective lives and that they will be socially transformative, the need to protect users' rights via effective governance has risen to the fore. This need becomes particularly compelling if we consider that misplaced trust in AI assistants or their developers may impact not just the individual user but also society at large, as a result of competitive scenarios between different AI assistants (see Section 6.2), or due to widening inequalities between 'haves' and 'have-nots' or between groups with differential access to AI assistants (see Section 5.2) [24]. Effective governance could reduce the power imbalances that exist between users and developers when the latter have unilateral ability to verify the trustworthiness of AI assistants.

However, the development and deployment of advanced AI assistants may rest on a complex system of base models, assistant applications and assistant tools. This gives rise to a final challenge in the field of trustworthy governance, namely the 'many hands' problem [19] which makes it difficult to create mechanisms that ensure that there are no accountability gaps and that roles are well-defined [4, 7, 12, 128]. More precisely, while base models are 'task agnostic,' AI assistants are a specific application of such models – to assist users by planning and executing sequences of actions on their behalf. In cases where something goes wrong, this raises the question of who should be considered morally accountable or liable – foundation model developers, who have control over the models but may struggle to anticipate every possible applications and associated risks, or AI assistant deployers, who do not necessarily have access to the underlying model [119].[13]

7 CONCLUSION AND PATHS AHEAD

This paper built on philosophical studies of trust and social sciences and policy work on human trust in AI to investigate when user trust in advanced AI assistants is justified. We argued that interactions between users and advanced AI assistants involve different targets of trust, namely AI assistants and their developers, and that understanding when users trust is justified requires considerations around both the competence and alignment of these different targets. By adopting a sociotechnical lens, we made recommendations about the interventions that should be implemented at the AI assistant design, organisational practices and third-party governance levels to help users discern when their trust in the competence and alignment of AI assistants and their developers is justified. The implementation of these interventions would reduce the risk of users becoming vulnerable to accidents that result from assistants or developers lacking the competence users expect them to have, or to misaligned values on the part of assistants or their developers.

The analysis presented here has certain limitations. First, the risks associated with misplaced user trust, across different dimensions, that we discussed in this paper only meant to serve as illustrative examples to enable us to introduce our account of justified trust in the context of user-AI assistant interactions. It is therefore important that future research is undertaken to develop a broader *taxonomy* of trust-related risks associated with advanced AI assistants to inform the responsible development and deployment of this technology. Second, because advanced AI assistants have yet to be deployed in the real world, the trust-related risks discussed in this paper are speculative, despite being grounded on existing empirical evidence of human trust in AI systems. Therefore, the considerations made here need to be complemented by *holistic evaluations* of AI assistants, particularly around trust-related risks of harm that may arise at the user-AI interaction and societal levels, so that effective mitigations can be implemented at the three levels. On this point, this paper focused primarily on AI assistants as the object of user trust. As AI assistants may become the main way in which users interface with other users, future studies should focus on assistants as the *mediator* of trust between users [134]. In addition, AI assistants may end up mediating the information users receive about the world, for example by summarising news articles in a way that is tailored to users' preferred communication style. Thus, research should explore the ways in which assistants may contribute to generalised trust or distrust at the societal level, for example through the spread of misinformation or by contributing to the creation of echo chambers. Third, the operationalisation of justified user trust in the competence and alignment of AI assistants and their developers leaves open many technical and normative questions that should become the focus of future investigations. These include questions around when evaluation of AI assistants should be considered sufficient to provide enough evidence for users to place justified trust in the technology [119]; or whose interests and values AI assistants should align with to be considered trustworthy, given that individual users' interests and values may be harmful to the self or others [33].

Advanced AI assistants may have important technological and societal ramifications. Thus, with this paper we aimed to raise awareness around the opportunities and risks associated with this emerging technology and the importance that user trust in assistants and their developers is justified. In this way, we hope to have inspired further research and policy work that will take advantage of the window of opportunity we are currently presented with to harness the transformative potential of AI assistants while ensuring their responsible and ethical development and deployment.

8 IMPACT STATEMENTS

8.1 Ethical Considerations Statement

This is a theoretical paper. It did not involve experiments with users and/or deployed systems, nor did it rely on sensitive user data.

8.2 Researcher Positionality Statement

Given the cultural and educational backgrounds of the authors, this paper builds on Western philosophical accounts of trust. Moreover, most policy frameworks on trustworthy AI originated from North American and European institutions.[14] This means that the considerations developed in this paper are unlikely to be representative of cultural differences in understandings of trust and trustworthiness around the world, and that authors with different backgrounds could have developed different recommendations around how to

[13] See this discussion playing out in the context of the EU AI Act [25], and proposed recommendations [36, 79].

[14] An exception is the China Academy for Information and Communication Technology's White Paper on Trustworthy Artificial Intelligence.

support users' justified trust. We therefore welcome additional perspectives to address possible limitations in our conceptualisation of user trust in AI assistants and their developers.

8.3 Adverse Impact Statement

Given the theoretical nature of the paper, we do not foresee any direct adverse impact of this work. However, we note that findings from empirical studies on when and how humans place trust in AI systems, which informs our philosophical account, could be misused by actors who have incentives to develop AI assistants with features that will induce users to trust them even when this is inappropriate. This risk highlights the importance of implementing the recommendations we make in Section 6.

ACKNOWLEDGMENTS

The authors would like to thank the FAccT reviewers for their helpful feedback, as well as Laura Weidinger and Meredith Ringel Morris for their detailed comments prior to submission. The authors also thank Edward Hughes, Pawan Mudigonda, Lewis Ho and Toby Shevlane for their early feedback and for inspiring some of the content of this paper.

REFERENCES

[1] Gavin Abercrombie, Amanda Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. On Anthropomorphism in Dialogue Systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 4776–4790. https://doi.org/10.18653/v1/2023.emnlp-main.290

[2] Lize Alberts and Max Van Kleek. 2023. Computers as Bad Social Actors: Dark Patterns and Anti-Patterns in Interfaces that Act Socially. arXiv:2302.04720 [cs.HC]

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. arXiv:1606.06565 [cs.AI]

[4] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf. 2023. Frontier AI Regulation: Managing Emerging Risks to Public Safety. arXiv:2307.03718 [cs.CY]

[5] Anthropic. 2023. Core Views on AI Safety: When, Why, What, and How. https://www.anthropic.com/index/core-views-on-ai-safety.

[6] Anthropic. 2023. Frontier Threats Red Teaming for AI Safety. https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety.

[7] Carolyn Ashurst. 2023. How to Regulate Foundation Models Can We Do Better than the EU AI Act? https://www.turing.ac.uk/research/interest-groups/fairness-transparency-privacy/how-to-regulate-foundation-models.

[8] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862 [cs.CL]

[9] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL]

[10] Natalie F Banner. 2020. The human side of health data. *Nature Medicine* 26, 7 (2020), 995–995.

[11] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 6, 8 pages. https://doi.org/10.1145/3551624.3555290

[12] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the Opportunities and Risks of Foundation Models. *CoRR* abs/2108.07258 (2021). arXiv:2108.07258 https://arxiv.org/abs/2108.07258

[13] Mark Bovens. 2014. Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism. In *Accountability and European governance*. Routledge, 18–39.

[14] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. arXiv:2004.07213 [cs.CY]

[15] Alignment Research Centre. 2023. Update on ARC's Recent Eval Efforts. https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/.

[16] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 651–666. https://doi.org/10.1145/3593013.3594033

[17] Avishek Choudhury and Hamid Shamszare. 2023. Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *Journal of Medical Internet Research* 25 (2023), e47184.

[18] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113.

[19] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1186–1197.

[20] Mark Coeckelbergh. 2012. Can we trust robots? *Ethics and information technology* 14 (2012), 53–60.

[21] Daniel C Dennett. 1989. *The intentional stance*. MIT press.

[22] Department for Science, Innovation and Technology. 2023. A Pro-Innovation Approach to AI Regulation. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

[23] S Devitt. 2018. Trustworthiness of autonomous systems. *Foundations of trusted autonomy (Studies in Systems, Decision and Control, Volume 117)* (2018), 161–184.

[24] Paul DiMaggio, Eszter Hargittai, et al. 2001. From the 'digital divide' to 'digital inequality': Studying Internet use as penetration increases. *Princeton: Center for Arts and Cultural Policy Studies, Woodrow Wilson School, Princeton University* 4, 1 (2001), 4–2.

[25] Connor Dunlop. 2023. An EU AI Act That Works for People and Society. https://www.adalovelaceinstitute.org/policy-briefing/eu-ai-act-trilogues/.

[26] Juan M Durán and Nico Formanek. 2018. Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines* 28 (2018), 645–666.

[27] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. arXiv:2303.10130 [econ.GN]

[28] European Commission. 2021. The EU AI Act. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[29] Tom Everitt, Ryan Carey, Eric D Langlois, Pedro A Ortega, and Shane Legg. 2021. Agent incentives: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11487–11495.

[30] Andrea Ferrario, Michele Loi, and Eleonora Viganò. 2020. In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology* 33 (2020), 523–539.

[31] Luciano Floridi and Josh Cowls. 2019. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review* 1, 1 (jul 1 2019). https://hdsr.mitpress.mit.edu/pub/l0jsh9d1.

[32] Ori Freiman. 2023. Making sense of the conceptual nonsense 'trustworthy AI'. *AI and Ethics* 3, 4 (2023), 1351–1360.

[33] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.

[34] Iason Gabriel and Vafa Ghazavi. 2023. The Challenge of Value Alignment: From Fairer Algorithms to AI Safety. In *Oxford Handbook of Digital Ethics*, Carissa Véliz (Ed.). Oxford University Press, Chapter 18, 336–355.

[35] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. 2024. The Ethics of Advanced AI Assistants. arXiv:2404.16244 [cs.CY]

[36] Maximilian Gahntz. 2023. The EU's AI Act and Foundation Models: The Final Stretch. https://foundation.mozilla.org/en/blog/the-eus-ai-act-and-foundation-models-the-final-stretch/.

[37] Andrew Gambino, Jesse Fox, and Rabindra A Ratan. 2020. Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication* 1 (2020), 71–85.

[38] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and Surprise in Large Generative Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1747–1764. https://doi.org/10.1145/3531146.3533229

[39] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.

[40] Google. 2023. AI Principles Progress Update 2023. https://ai.google/static/documents/ai-principles-2023-progress-update.pdf.

[41] Google. 2023. Assistant with Bard: A step toward a more personal assistant. https://blog.google/products/assistant/google-assistant-bard-generative-ai/.

[42] Google DeepMind. 2022. How Our Principles Helped Define AlphaFold's Release. https://deepmind.google/discover/blog/how-our-principles-helped-define-alphafolds-release/.

[43] Google DeepMind. 2023. Transforming The Future of Music Creation. https://deepmind.google/discover/blog/transforming-the-future-of-music-creation/.

[44] Mackenzie Graham. 2021. Data for sale: trust, confidence and sharing health data with commercial companies. *Journal of Medical Ethics* (2021).

[45] Mackenzie Graham, Richard Milne, Paige Fitzsimmons, and Mark Sheehan. 2022. Trust and the Goldacre Review: Why trusted research environments are not about trust. *Journal of Medical Ethics* (2022).

[46] Katherine Hawley. 2014. Trust, Distrust and Commitment. *Noûs* 48, 1 (2014), 1–20. https://doi.org/10.1111/nous.12000 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/nous.12000

[47] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2022. Unsolved Problems in ML Safety. arXiv:2109.13916 [cs.LG]

[48] High-Level Expert Group on Artificial Intelligence. 2019. Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

[49] Russell Hotten. 2015. Volkswagen: The Scandal Explained. https://www.bbc.co.uk/news/business-34324772.

[50] Seaton Huang, Helen Toner, Zac Haluza, Rogier Creemers, and Graham Webster. 2023. Translation: Measures for the Management of Generative Artificial Intelligence Services (Draft for Comment) – April 2023. https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/.

[51] Inflection. 2023. I'm Pi, Your personal AI. https://inflection.ai/.

[52] Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. arXiv:1805.00899 [stat.ML]

[53] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. 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* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 624–635. https://doi.org/10.1145/3442188.3445923

[54] Karen Jones. 1996. Trust as an Affective Attitude. *Ethics* 107, 1 (1996), 4–25. https://doi.org/10.1086/233694

[55] Eirini Kalliamvakou. 2022. Research: Quantifying GitHub Copilot's Impact on Developer Productivity and Happiness. https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/.

[56] Alexandra D Kaplan, Theresa T Kessler, J Christopher Brill, and PA Hancock. 2023. Trust in artificial intelligence: Meta-analytic findings. *Human factors* 65, 2 (2023), 337–359.

[57] Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. 2022. Discovering Agents. arXiv:2208.08345 [cs.AI]

[58] Angeliki Kerasidou. 2017. Trust me, I'ma researcher!: The role of trust in biomedical research. *Medicine, Health Care and Philosophy* 20, 1 (2017), 43–50.

[59] Charalampia Xaroula Kerasidou, Angeliki Kerasidou, Monika Buscher, and Stephen Wilkinson. 2022. Before and beyond trust: Reliance in medical AI. *Journal of medical ethics* (2022).

[60] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 77–88. https://doi.org/10.1145/3593013.3593978

[61] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. 2020. Specification Gaming: The Flip Side of AI Ingenuity. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

[62] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2022. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society* (2022), 14614448221142007.

[63] Nancy K Lankton, D Harrison McKnight, and John Tripp. 2015. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems* 16, 10 (2015), 1.

[64] Johann Laux, Sandra Wachter, and Brent Mittelstadt. 2023. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance* (2023).

[65] Seth Lazar and Alondra Nelson. 2023. AI safety on whose terms? *Science* 381, 6654 (2023), 138–138. https://doi.org/10.1126/science.adi8982

[66] Q.Vera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (<conf-loc>, <city>Seoul</city>, <country>Republic of Korea</country>, </conf-loc>) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1257–1268. https://doi.org/10.1145/3531146.3533182

[67] Alexandra Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2023. Power Hungry Processing: Watts Driving the Cost of AI Deployment? arXiv:2311.16863 [cs.LG]

[68] Stephen Marsh, Tosan Atele-Williams, Anirban Basu, Natasha Dwyer, Peter R Lewis, Hector Miller-Bakewell, and Jeremy Pitt. 2020. Thinking about trust: People, process, and place. *Patterns* 1, 3 (2020).

[69] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[70] Daniel J McAllister. 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal* 38, 1 (1995), 24–59.

[71] Kevin R. McKee, Xuechunzi Bai, and Susan T. Fiske. 2023. Humans perceive warmth and competence in artificial intelligence. *iScience* 26, 8 (2023), 107256. https://doi.org/10.1016/j.isci.2023.107256

[72] Meta. 2023. Introducing New AI Experiences Across Our Family of Apps and Devices. https://about.fb.com/news/2023/09/introducing-ai-powered-assistants-characters-and-creative-tools/.

[73] Jacob Metcalf, Ranjit Singh, Emanuel Moss, Emnet Tafesse, and Elizabeth Anne Watkins. 2023. Taking Algorithms to Courts: A Relational Approach to Algorithmic Accountability. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1450–1462. https://doi.org/10.1145/3593013.3594092

[74] Microsoft. 2023. Copilot: Your everyday companion. https://copilot.microsoft.com/.

[75] Wei Peng Minjin Rheu, Ji Youn Shin and Jina Huh-Yoo. 2021. Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design. *International Journal of Human–Computer Interaction* 37, 1 (2021), 81–96. https://doi.org/10.1080/10447318.2020.1807710 arXiv:https://doi.org/10.1080/10447318.2020.1807710

[76] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[77] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature machine intelligence* 1, 11 (2019), 501–507.

[78] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: A three-layered approach. *AI and Ethics* (2023), 1–31.

[79] Sarah Myers West. 2023. General Purpose AI Poses Serious Risks, Should Not Be Excluded From the EU's AI Act | Policy Brief. https://ainowinstitute.org/publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act.

[80] Arvind Narayanan. 2019. How to Recognize AI Snake Oil. https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf.

[81] Jessica Newman. 2023. A Taxonomy of Trustworthiness for Artificial Intelligence: Connecting Properties of Trustworthiness with Risk Management and the AI Lifecycle. https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/.

[82] C. Thi Nguyen. 2022. Transparency is Surveillance. *Philosophy and Phenomenological Research* 105, 2 (2022), 331–361. https://doi.org/10.1111/phpr.12823 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/phpr.12823

[83] Philip J Nickel, Maarten Franssen, and Peter Kroes. 2010. Can we make sense of the notion of trustworthy technology? *Knowledge, Technology & Policy* 23 (2010), 429–444.

[84] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and engineering ethics* 2 (1996), 25–42.

[85] NIST. 2023. Artificial Intelligence Risk Management Framework. https://doi.org/10.6028/nist.ai.100-1.

[86] Benjamin Noah and Arathi Sethumadhavan. 2019. Generational differences in trust in digital assistants. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 206–210.

[87] OECD. 2021. Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems. https://www.oecd.org/science/tools-for-trustworthy-ai-008232ec-en.htm.

[88] Office of Science and Technology Policy. 2022. Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[89] Mancur Olson. 1965. *The Logic of Collective Action*. Harvard University Press.

[90] OpenAI. 2019. GPT-4 System Card. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

[91] OpenAI. 2023. Assistants API. https://platform.openai.com/docs/assistants/overview.

[92] Onora O'Neill. 2002. *Autonomy and Trust in Bioethics*. Cambridge University Press.

[93] Onora O'Neill. 2018. Linking Trust to Trustworthiness. *International Journal of Philosophical Studies* 26, 2 (2018), 293–300.

[94] Onora O'Neill and James Bardrick. 2017. Trust, Trustworthiness And Transparency. https://www.thebritishacademy.ac.uk/documents/2563/Future-of-the-corporation-Trust-trustworthiness-transparency.pdf.

[95] Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. 2023. ART: Automatic multi-step reasoning and tool-use for large language models. arXiv:2303.09014 [cs.CL]

[96] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2023. AI Deception: A Survey of Examples, Risks, and Potential Solutions. arXiv:2308.14752 [cs.CY]

[97] Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv:2302.06590 [cs.SE]

[98] Valentina Pitardi and Hannah R Marriott. 2021. Alexa, she's not human but… Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology & Marketing* 38, 4 (2021), 626–642.

[99] Joseph C Pitt. 2010. It's not about technology. *Knowledge, Technology & Policy* 23 (2010), 445–454.

[100] Eric Racine, Tristana Martin Rubio, Jennifer Chandler, Cynthia Forlini, and Jayne Lucke. 2014. The value and pitfalls of speculation about science and technology in bioethics: the case of cognitive enhancement. *Medicine, Health Care and Philosophy* 17 (2014), 325–337.

[101] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 959–972. https://doi.org/10.1145/3531146.3533158

[102] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 33–44. https://doi.org/10.1145/3351095.3372873

[103] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 557–571. https://doi.org/10.1145/3514094.3534181

[104] Anatol Rapoport and Albert M Chammah. 1966. The game of chicken. *American Behavioral Scientist* 10, 3 (1966), 10–28.

[105] Karoline Reinhardt. 2023. Trust and trustworthiness in AI ethics. *AI and Ethics* 3, 3 (2023), 735–744.

[106] Gernot Rieder, Judith Simon, and Pak-Hang Wong. 2021. Mapping the Stony Road toward Trustworthy AI. *Machines we trust: Perspectives on dependable AI* (2021), 27.

[107] Alan Rubel, Adam Pham, and Clinton Castro. 2019. Agency Laundering and Algorithmic Decision Systems. In *Information in Contemporary Society: 14th International Conference, iConference 2019, Washington, DC, USA, March 31–April 3, 2019, Proceedings 14*. Springer, 590–598.

[108] Mark Ryan. 2020. In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics* 26, 5 (2020), 2749–2767.

[109] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks. arXiv:2207.13243 [cs.LG]

[110] Lucile Saulnier, Siddharth Karamcheti, Hugo Laurençon, Leo Tronchon, Thomas Wang, Victor Sanh, and Amanpreet Singh. 2022. Putting Ethical Principles at the Core of the Research Lifecycle. https://huggingface.co/blog/ethical-charter-multimodal.

[111] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are Emergent Abilities of Large Language Models a Mirage? arXiv:2304.15004 [cs.AI]

[112] Nicolas Scharowski, Michaela Benk, Swen J. Kühne, Léane Wettstein, and Florian Brühlmann. 2023. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 248–260. https://doi.org/10.1145/3593013.3593994

[113] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. arXiv:2302.04761 [cs.CL]

[114] Donna Schreuter, Peter van der Putten, and Maarten H Lamers. 2021. Trust me on this one: Conforming to conversational assistants. *Minds and Machines* 31 (2021), 535–562.

[115] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[116] Hetan Shah. 2018. Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2128 (2018), 20170362.

[117] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. arXiv:2210.01790 [cs.LG]

[118] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature* 623, 7987 (2023), 493–498.

[119] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul McMillan, Alex Beutel, Alexandre Passos, and David G. Robinson. 2023. Practices for Governing Agentic AI Systems. https://openai.com/research/practices-for-governing-agentic-ai-systems.

[120] Mark Sheehan, Phoebe Friesen, Adrian Balmer, Corina Cheeks, Sara Davidson, James Devereux, Douglas Findlay, Katharine Keats-Rohan, Rob Lawrence, and Kamran Shafiq. 2021. Trust, Trustworthiness and Sharing Patient Data for Research. *Journal of Medical Ethics* 47, 12 (2021), 26–26. https://doi.org/10.1136/medethics-2019-106048

[121] Stephanie Sheir, Arianna Manzini, Helen Smith, and Jonathan CS Ives. 2024. Adaptable Robots, Ethics, and Trust: A Qualitative and Philosophical Exploration of the Individual Experience of Trustworthy AI. *AI and Society* (2024).

[122] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. 2023. Model evaluation for extreme risks. arXiv:2305.15324 [cs.AI]

[123] Mona Simion and Christoph Kelp. 2023. Trustworthy artificial intelligence. *Asian Journal of Philosophy* 1, 2 (March 3 2023). https://link.springer.com/article/10.1007/s44204-023-00063-5.

[124] Sangeeta Singh-Kurtz. 2023. The Man of Your Dreams. https://www.thecut.com/article/ai-artificial-intelligence-chatbot-replika-boyfriend.html.

[125] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies* 168 (2022), 102903.

[126] Helen Smith, Arianna Manzini, Mari-Rose Kennedy, and Jonathan Ives. 2023. Ethics of Trust/Worthiness in Autonomous Systems: A Scoping Review.. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems* (Edinburgh, United Kingdom) *(TAS '23)*. Association for Computing Machinery, New York, NY, USA, Article 20, 15 pages. https://doi.org/10.1145/3597512.3600207

[127] Irene Solaiman. 2023. The Gradient of Generative AI Release: Methods and Considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 111–122. https://doi.org/10.1145/3593013.3593981

[128] Merlin Stein and Connor Dunlop. 2023. Safe Before Sale. https://www.adalovelaceinstitute.org/report/safe-before-sale/.

[129] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2013. Developing a framework for responsible innovation. *Research Policy* 42, 9 (2013), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

[130] The Adaptive Agents Group. 2021. The Shibboleth Rule for Artificial Agents. https://hai.stanford.edu/news/shibboleth-rule-artificial-agents.

[131] The White House. 2023. Ensuring Safe, Secure, and Trustworthy AI. https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

[132] The White House. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[133] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2021. Trustworthy artificial intelligence. *Electronic Markets* 31 (2021), 447–464.

[134] Lauren Thornton, Bran Knowles, and Gordon Blair. 2022. The Alchemy of Trust: The Creative Act of Designing Trustworthy Socio-Technical Systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1387–1398. https://doi.org/10.1145/3531146.3533196

[135] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 272–283. https://doi.org/10.1145/3351095.3372834

[136] Andrew Trask, Emma Bluemke, Ben Garfinkel, Claudia Ghezzou Cuervas-Mons, and Allan Dafoe. 2020. Beyond Privacy Trade-offs with Structured Transparency. arXiv:2012.08347 [cs.CR]

[137] US Department of Justice. 2022. Court Issues Order Requiring Cigarette Companies to Post Corrective Statements; Resolves Historic RICO Tobacco Litigation. https://www.justice.gov/opa/pr/court-issues-order-requiring-cigarette-companies-post-corrective-statements-resolves-historic.

[138] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2024. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. arXiv:2306.11698 [cs.CL]

[139] Francis Rhys Ward, Francesco Belardinelli, Francesca Toni, and Tom Everitt. 2023. Honesty Is the Best Policy: Defining and Mitigating AI Deception. arXiv:2312.01350 [cs.AI]

[140] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. arXiv:2206.07682 [cs.CL]

[141] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986 [cs.AI]

[142] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks Posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088

[143] Jess Whittlestone and Jack Clark. 2021. Why and How Governments Should Monitor AI Development. arXiv:2108.12427 [cs.CY]

[144] Maranke Wieringa. 2020. What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3351095.3372833

[145] Jan W Wiktor and Katarzyna Sanak-Kosmowska. 2021. *Information asymmetry in online advertising*. Routledge.

[146] Naim Zierau, Christian Engel, Matthias Söllner, and Jan Marco Leimeister. 2020. Trust in smart personal assistants: A systematic literature review and development of a research agenda. In *International Conference on Wirtschaftsinformatik (WI).-Potsdam, Germany*.