Making Sense of the Conceptual Nonsense 'Trustworthy AI'

Ori Freiman (Ethics of AI Lab, University of Toronto)

Following the publication of numerous ethical principles and guidelines, the concept of 'Trustworthy AI' has become widely used. However, several AI ethicists argue against using this concept, often backing their arguments with decades of conceptual analyses made by scholars who studied the concept of trust. In this paper, I explain the historical-philosophical roots of their objection and the premise that trust entails a human quality that technologies lack. Then, I review existing criticisms about 'Trustworthy AI' and the consequence of ignoring these criticisms: if the concept of 'Trustworthy AI' is kept being used, we risk attributing responsibilities to agents who cannot be held responsible, and consequently, deteriorate social structures which regard accountability and liability. Nevertheless, despite suggestions to shift the paradigm from 'Trustworthy AI' to 'Reliable AI', I argue that, realistically, this concept will be kept being used. I end by arguing that, ultimately, AI ethics is also about power, social justice, and scholarly activism. Therefore, I propose that community-driven and social justice-oriented ethicists of AI and trust scholars further focus on (a) democratic aspects of trust formation; and (b) draw attention to critical social aspects highlighted by phenomena of distrust. This way, it will be possible to further reveal shifts in power relations, challenge unfair status quos, and suggest meaningful ways to keep the interests of citizens.

Keywords: Trustworthy AI; Ethics of AI; Trust

1. Introduction

AI policy documents that include ethical principles and guidelines have mushroomed in recent years and many of these documents include the concept of 'Trustworthy AI' (Schiff et al. 2021; Floridi & Cowls 2021; Jobin, Ienca, & Vayena 2019; Hagendorff 2020; Fjeld 2020; c.f. Dotan 2021; Renda 2020; Mittelstadt 2019). As time goes, this concept shifted from the preliminary principles and guidelines to actual proposed legislation and regulation, and other initiatives. For example, the US (NAII 2021), China (CAICT 2021), and the European Union (EC 2021) advance legislation by building on this concept. By now, the concept of 'Trustworthy AI' has become widely used among policymakers, technologists, investors, and scholars from diverse fields.

One of the notable initiatives that have let to the popularization of this concept was the European Commission's (EC) High-Level Expert Group (HLEG) on AI. Their first draft was published in

December 2018 (HLEG 2018) and sparked interest in this term[1]. The HLEG's document, *Ethics Guidelines for Trustworthy AI* (2019), received mixed reactions. They ranged from positive responses to constructively pointing out critical aspects (Rieder, Simon, & Wong 2021). Among the responses, some questioned whether these principles actually mean anything and commented that "AI regulation needs action not philosophical thinking" (Davies 2019). Others have criticized the processes and procedures of HLEG, such as producing this document while being pressured to deliver results - without sufficient time for deliberation (Gießler et al. 2019; AlgorithmWatch 2019). Arguably, the stingiest criticism was that the HLEG was excessively influenced by industry interests "at the cost of civil society's concerns, and for being an exercise in ethics washing" (Article 19 2019).

The issue of ethics-washing in AI ethics poses a series of concerns. Roughly, ethics-washing is "embracing the language of ethics to defuse criticism and resist government regulation, without committing to ethical behavior", and "a façade of ethics while largely continuing with business-as-usual" (Green 2022). Abusing and misusing ethics can prevent the implementation of necessary regulations (Rességuier & Rodrigues 2020).

A common tactic for ethics-washing is to keep ethical debates running while the technology is developed to delay or avoid strict regulations and minimize the regulatory effect (Peukert & Kloker 2020). Floridi (2019) identifies other tactics, such as imposing self-regulation while lobbying against new legal norms ("ethics-lobbying"); engaging only with particular ethical principles that retrospectively justify a company's behaviour ("ethics-shopping"); employing PR activity to address an ethical problem ("ethics-bluewashing"); and outsourcing non-ethical training of algorithms to countries will less regulation ("ethics-dumping").

In the context of HLEG, the most scathing criticism was arguably about ethics-washing. This charge came from philosophers and ethicists of AI Thomas Metzinger and Mark Coeckelbergh. Not only both are well-established scholars in their fields, but also both are members of the HLEG, which renders the accusation as much worse. They expressed disappointment and dissatisfaction with the opportunity to affect the outcome of HLEG: "We sadly conclude that the few ethicists among the 52 members of the AI advisory group have been nothing but a fig leaf" (Metzinger & Coeckelbergh 2020). According to Metzinger (2019) "[t]he Trustworthy AI story is a marketing narrative invented by industry, a bedtime story for tomorrow's customers", and the "Trustworthy AI narrative is, in reality, about developing future markets and using ethics debates as elegant public decorations for a large-scale investment strategy"[2].

Additionally, fundamental to Metzinger's criticism of 'Trustworthy AI' as a narrative constructed by the HLEG, stands his examination of 'Trustworthy AI' as a concept:

> The underlying guiding idea of a "trustworthy AI" is, first and foremost, *conceptual nonsense***.** Machines are not trustworthy; only humans can be trustworthy (or untrustworthy). If, in the future, an untrustworthy corporation or government behaves unethically and possesses

[1] While I could not trace the source of the term 'Trustworthy AI', the initial popularization of it can be associated to the HLEG's first draft. See Google Trends worldwide search for "Trustworthy AI".

[2] Despite the criticism aimed at HLEG, Metzinger (2019) acknowledged that this initiative is "currently the best globally available platform for the next phase of discussion".

> good, robust AI technology, this will enable more effective unethical behaviour. (2019, emphasis added)

With these claims, it is possible to analytically separate the criticism of HLEG and the "Trustworthy AI" narrative from criticism about the usage of the concept of 'Trustworthy AI'. This paper focuses on the latter.

2. Anthropocentric View of Trust

At the foundations of the concept of 'Trustworthy AI' stands the concept of trust. The concept of trust is ubiquitous, and local applications and understandings of the concept occur wherever it is mentioned. For example, for economists, trust is usually discussed in terms of calculation and risk assessment, and for sociologists, trust is the enabler of social interaction (Keymolen 2016: 13). After surveying much of the literature on trust in philosophy, Judith Simon (2013) notes that "as pervasive trust appears as a phenomenon, as elusive it seems as a concept". After all, trust in the context of democracy and political philosophy is different from trust in Science or trust that someone else's say so is true.

Trust scholars, philosophers of technology and ethicists of AI oppose to the term 'Trustworthy AI'. To understand this opposition, I turn to explain the theoretical roots of the concept of trust, and the incompatibility to associate it with technologies. These roots are found within social epistemology - a subfield of philosophy that deals with the social dimensions of knowledge.

Social epistemology is closely tied with several other subfields, such as Science and Technology Studies (STS) and feminist epistemology. The field's roots are found within traditional Anglo-American analytic philosophy. Within social epistemology, the standard view on trust is that trust relations are based on a human quality such as goodwill. Therefore, trust relations are only possible between two agents. These agents are individual persons, however, on a generous interpretation, they also involve groups (Freiman 2014; Miller & Freiman 2020).

This view rests upon a commonly acknowledged distinction between a genuine trust and mere reliance (Hawley 2014; Tallant 2019). The source of this distinction is often traced to the three ideas: (a) "trusting can be betrayed, or at least let down, and not just disappointed" (Baier 1986: 235); (b) trusting is "inherently subject to the risk that the other will abuse the power of discretion" (Hardin 1993: 507); and that therefore, (c) "trusting is not an attitude that we can adopt toward machinery" (Jones 1996: 14). Unlike genuine trust, mere reliance is a way of acting in light of the probability that technology will perform successfully (Nickel 2013). Genuine trust entails a moral aspect, that mere reliance does not.

A related term to reliance is *reliability*. Reliability can be thought of as a law-like regularity, that can be predicted in calculations and discussed in terms of accuracy. Unlike reliability, trust is viewed as a kind of moral relationship. It ascribes a human quality, such as goodwill to the trustee (Baier 1986: 242, 252; Jones 1996: 14), or requires the moral obligation and dependence of the trustor on the trustee's responsiveness to fulfill their commitments (Jones 2012; Hawley 2014)[3]. The latter are

[3] For the relations of the concepts of trust and reliability to the concepts of confidence, risk, and vulnerability, see De Filippi, Mannan, & Reijers 2020: §2.3. and references within.

referred to as *responsiveness theories of trust* (Nguyen forthcoming). According to these theories, to trust someone is to hold a belief that the trustee will respond to the trust. It renders a trustworthy person as someone who "takes the fact that they are counted on to be a reason for acting as counted on" (Jones 2012: 66).

Overall, the attitude of trust entails an expectation for the trustee to fulfill their commitments and be aware that they are trusted. Failing to fulfill the trustor's expectations, or discovering that there is no such goodwill, can lead to a sense of betrayal. A trustworthy agent, therefore, has the power to betray the trustor. Can (trustworthy) AIs betray humans?

The field of social epistemology is "infused with anthropocentric concepts" (Humphreys 2009: 221), such as 'knowledge', 'testimony', and relevant to our case here - 'trust'. On the one hand, this approach seems outdated in the era of AIs. But, on the other hand, maintaining this approach might be necessary for the era of rapidly accelerating, pervasive technological development.

Currently, some offer a way around the impossibility to direct trust in technologies. Some suggest that "trust in technology" is a metaphorical and linguistic expression, actually referring to trusting the humans behind the technologies. This view was defended by Joseph Pitt, who termed the catchphrase "technology is *humanity* at work" (2010: 445, emphasis in original). They reduce issues of trust in technologies to issues of trust in the people and institutions who relate to the technologies, such as engineers and designers. For example, when a person crosses a bridge and trusts it not to collapse, she trusts its builders and those responsible for its maintenance (Origgi 2008).

This view describes several different approaches to trust, all commonly share two features: (a) not ascribing human agency to technologies; and (b) referring to humans behind the technologies in matters regarding trust and responsibility. Call these approaches *The Anthropocentric View of Trust.* This view has also been labelled the 'reductive view' (Nickel 2022), 'humans behind the machines' (Freiman 2021), 'indirect trust' (Coeckelbergh 2012), and 'human-centered terminology from philosophical accounts' (Rieder, Simon & Wong 2021). They all commonly point out that the concept of trust, according to the traditional conception, is not suitable for technological artifacts (Freiman & Miller 2020; Freiman 2021)[4].

Not surprisingly, the theoretical incompatibility of 'trust' and AI did not escape scholars. For example, Weydner-Volkmann & Feiten (2021) defend the concept of 'trust in technology' to make sense of the concept of 'Trustworthy AI' by opposing the traditional philosophical view of trust. They argue that the concept of 'trust in technology' enables addressing critical societal needs, and that the notion of 'trustworthy technology' enables us highlighting issues about laypeople who are vulnerable due to attacks or misuse of technologies.

Similarly, Braun, Bleher & Hummel (2021) oppose the anthropocentric view of trust and advocate against the skepticism that the notion of 'trustworthy AI' can bring. They interpret the traditional philosophical view as a narrow account of trust. As counterexamples against the narrow view, they

[4] It is possible to distinguish between strong and weak versions of the anthropocentric view. According to a strong version, technologies cannot be the actual objects of trust, and people and institutions are. On a weaker version, technologies are *ipso facto* objects of trust, yet it is only on a closer inspection that we recognize humans and institutions as additional objects of trust. Both versions reduce issues of trust in technologies to the people behind them and do not ascribe human agency to technologies.

discuss trust trusting a prosthesis, a guiding dog, and institutions – all non-humans. Ultimately, they call the skeptics about 'Trustworthy AI' not to oppose the project.

In various fields, such as machine ethics, information ethics, and parts of STS, it is accepted to attribute morality - and the ability to trust and know, to technologies (Freiman 2014). As a result, scholars within these fields research practical aspects of what it means to trust in AI (see Glikson & Woolley 2020 and references within). Similarly, Weydner-Volkmann & Feiten (2021) and Braun, Bleher & Hummel (2021) aligned with these views. However, other scholars have criticized the concept of 'trustworthy AI'.

3. Criticizing the Concept of 'Trustworthy AI'

In this section, I survey four criticisms of the concept of 'Trustworthy AI'. These are (i) Joanna Bryson's (2018) seminal arguments that AIs are not our peers and that it is possible to trace accountability back to humans; (ii) Margit Sutrop's (2019) argument about social cognition; (iii) Mark Ryan's (2020) argument about goodwill and the ability to normatively commit; and (iv) Gernot Rieder, Judith Simon, and Pak-Hang Wong's (2020) focus on the democratic and political aspects of 'Trustworthy AI'. I conclude this section by spelling what we risk by recognizing an AI as trustworthy –anthropomorphizing AIs and consequently falsely attributing them with responsibilities.

3.1. Peers and Accountability

Bryson (2018) demonstrates both versions of the anthropocentric view of trust. She has a dual argument against trusting AI: First, she argues that trust relations can only exist among peers - which AIs and humans are not. Second, she goes against common claims that accountability is impossible to trace and maintain since this is the nature of machine learning, the system is complex, or it is autonomous. Her second argument is that if accountability can be traced back to humans, there is no reason to trust AIs.

Underlying both of her arguments is that humans are running institutions, which are complex and autonomous, but still accountable. In many cases, logs of who does what and why are maintained. She gives the automotive industry as an example of a regulated industry that is held accountable for its products: when accidents happen, the companies that build autonomous cars are held accountable. She argues that the same logic should be applied to companies that develop software - intelligent or not. Her ultimate conclusion is that instead of trust in AI, we should discuss trust in humans: "When a system using AI causes damage, we need to know we can hold the human beings behind that system to account" (ibid).

3.2. Reliance and a One-Way Communication Process

Sutrop (2019) builds upon Bryson's analysis. She criticizes HLEG's guidelines for ignoring the philosophical literature's distinction between trust and reliance. Building upon this literature, she argues that trusting is a two-way communication process. In this process, the trustee expresses competence, goodwill, and commitment to what is expected, while the trustor has a desire that the trustee would be trustworthy.

Under such an analysis, her argument goes, trust is a social relation that involves social cognition. Therefore, it is possible to speak of trust in AI only in two cases: "when we speak about human-like AI or when we mean the individuals and institutions behind AI systems" (512). For Sutrop, direct trust in AI is possible. It just means trusting either a human-like AI or those behind the system.

Sutrop points out that many objects of trust and various trust relations exist: individuals or institutions who relate to design, manufacture, ownership, are responsible for its regulation, oversee its usage, etc. However, when we speak of AI designed to fulfil specific tasks, we rely on it to function well; and that reliance is a one-way communication process. Therefore, "we should better describe our attitude towards it as reliance" (ibid).

3.3. Goodwill and a Normative Commitment

Perhaps the most encompassing conceptual analysis of the concept of trust in the context of 'Trustworthy AI' comes from Ryan (2020). If Sutrop (2019) calls to change the description of our attitudes from trust to reliance, Ryan calls, laud and clear, to replace the 'Trustworthy AI' paradigm with the 'Reliable AI' approach.

Ryan (2020) argues that for the HLEG, the object of trust can be (i) the AI technology itself; (ii) the people and organizations behind the AI; and (iii) the socio-technical systems as a whole. However, he contends, trust cannot be directed at AI, since it is problematic to associate human moral activities with AI. To support his view on trust, he differentiates between affective, normative, and rational accounts of trust in the philosophical literature.

An *affective account of trust* is characterized by having the trustor believe in the trustee's goodwill and expect that the trustee will be motivated by this affection. The fact is that AI is not able to be motivated by goodwill and the expectation that someone is counting on it.

In the *normative account of trust,* the trustor's expectations are directed not only on what the trustee will do, but also on what they *should* do. From the side of the trustee, it entails a motivation to be morally responsible for their actions - a normative commitment. In this case, too, AI is not capable of being morally responsible for its actions[5].

Lastly, in the *rational account of trust,* the trustor rationally calculates whether to trust the trustee. This account of trust is simply a matter of a one-sided prediction rather than the consequence of a two-way relationship between the trustor and trustee. Unlike the previous accounts, it does not require goodwill or a normative commitment, only mere calculations. Therefore, this account of trust "should not be called trust at all, as it is a form of reliance" (Ryan 2020).

3.4. Democratic and Political Culture

Ryan's (2020) critical engagement focused on conceptual analysis. Likewise, Rieder, Simon & Wong (2020) examine the usage of the concept of 'Trustworthy AI'. Underlying their examination is a

[5] For the topic of AI and assigning moral status, a degree of moral consideration, or moral agency, see, Gunkel 2018a, 2018b; Tavani 2018; Stamboliev 2020; Coeckelbergh 2020; Nyholm 2020; Danaher 2020; Farina 2022.

rejection of the rational account of trust, and adoption of an account that attributes motivations to the trustee. Their account is compatible with the affective and normative accounts of trust mentioned previously. Additionally, their view of trust acknowledges the anthropocentricity of 'trust', yet reduces it to : "trust in AI systems is plausible only to the extent that we include human agents as the targets of trust, thereby framing AI systems as socio-technical systems that include human agents".

They identify four epistemic and moral difficulties for attaining 'Trustworthy AI': the AI system must be (i) capable of fulfilling the trustor's expectation by being reliable; (ii) self-assess and monitor the system's limitations and promises to those who trust it; (iii) account for the interests and values of those who trust it (focusing on the most vulnerable groups of users); and (iv) require information disclosure about the inner working of the system and its goals, including financial interests. Rieder, Simon & Wong (2020) show how each of these difficulties can be met, therefore allowing, in principle, achieving a 'Trustworthy AI'.

However, they are skeptical about "recent efforts to sell trustworthy AI as a ready-made label or brand" (ibid). The moral requirements of their account of trust can only be met by cultivating a 'trustworthy AI culture'. In such a culture, the public is involved. This involvement is an alternative to mere standardization processes that can hardly satisfy moral requirements. Additionally, efforts to achieve safe and reliable products should not amount to trustworthiness: "trust should ultimately only be extended to the democratic and political culture surrounding these products" (ibid).

3.5. The Risk of Anthropomorphizing AI-Based Technologies and AI Systems

So far, I have explored four different anthropocentric views of trust. In this part, I ask what will happen if, to the discontentment of philosophers and AI ethicists - scholars, regulators, and the rest of the world will continue to characterize AI as trustworthy? As Ryan points out, we risk anthropomorphizing AIs. However, what exactly the risk of anthropomorphizing means?

Anthropomorphizing is ascribing human-like features and characteristics to an otherwise non-human object (Kontogiorgos et al. 2019) or referring "to the manner of attributing human reasoning to non-human beings" (Tamir & Zohar 1991: 57). When we anthropomorphize AI-based technologies, such as robots or digital assistants like Siri or Alexa, we often assign them moral status and a degree of moral consideration (see FN 5). In this case, these technologies are perceived as individual agents. However, anthropomorphism can also occur on a greater scale. For example, when we assign an AI system with human-like (or institution-like) characteristics, we risk falsely attributing it with a moral character it is incapable of having.

In both forms of anthropomorphism – individual and social, the risk is trusting that the object can be a moral agent. It is misplacing trust, i.e. directing trust at the wrong agent. When we misplace trust, we falsely attribute responsibilities to non-moral agents, social structures which regard liability, accountability and responsibility are altered. Institutions and the people behind the technologies – i.e. those designing, developing, employing, auditing, and maintaining the AI systems, are not the objects of proper moral and legal scrutiny.

4. Trust Scholars and AI Ethicists: What's Next?

Overall, following Bryson (2018), the criticisms mentioned above seem twofold and interconnected: (a) denying that we can meaningfully directly trust AIs; and (b) that we should not trust AI. Genuine trust entails a human quality and depends on affective, normative and moral attitudes that technologies, at least currently, cannot have. Philosophers and AI ethicists are left with the question of how to build upon these criticisms and have real-world practical influence.

4.1. The Shift from 'Trustworthy AI' to 'Reliable AI' will (Probably) Not Happen

There is a conceptual gap regarding the usage of 'Trustworthy AI'. While some use the concept freely, others restrict its usage. The latter represents primarily philosophers and ethicists of AI, who are aware of the anthropocentricity of the concept of 'trustworthy' and hold an anthropocentric view of trust. The concept 'Trustworthy AI' has been labelled as a "conceptual nonsense" (Metzinger 2019), "conceptual misunderstanding" (Hatherley 2020), and "a misnomer" (Braun, Bleher, & Hummel 2021).

In some cases, there might be a normative expectation of those who hold the anthropocentric view of trust that others will subscribe to the restrictive usage of the concept. Failing to do so might lead "to a corrupted form of trust in a domain where rightful trust is of paramount importance" (Nickel 2022). Often, scholars who analyze the term suggest replacing 'Trustworthy AI' with 'Reliable AI'. Doing so would enable to avoid associating AIs as trustworthy and focus on the correct objects of trust: institutions and people behind AIs. For example, Joshua James Hatherley argues:

> Rather than trustworthy AI, this pursuit may be better served by being reframed in terms of reliable AI, reserving the label of 'trust' for reciprocal relations between beings with agency (Hatherley 2020).

Ryan (2020), too, concludes with a similar call to replace the concept of 'Trustworthy AI' with a 'Reliable AI Approach'. The reason is to ensure that we focus on the important domains of trust:

> Overall, proponents of AI ethics should abandon the 'trustworthy AI' paradigm as it is too fraught with problems, replacing it with the reliable AI approach, instead. (Ryan 2020: 17)

Ryan's call to replace the terminology is directed at proponents of AI ethics. Within the scholarly field of AI ethics, there are normative expectations to subscribe to the anthropocentric view of trust and therefore use the reliable AI approach. After all, sharing similar epistemic norms and standards and having similar social conventions enable peers within the same discipline to trust each other (Wilholt 2009).

However, expectations from AI ethics proponents, be they scholars from different disciplines, policymakers, technologists, investors, or others, will not join this change. It seems that the larger AI ethics community is too big to follow this change, and there are many other pressing issues to deal with (see, e.g., Dubber, Pasquale & Das 2020; Vesnic-Alujevic, Nascimento, & Polvora 2020).

Trying to correct others who are unaware of the philosophical underpinning of the concept of 'trust', arguably amounts to a waste of limited resources. It is reasonable to say that such efforts would most likely be ignored, and even if taken seriously, some would still defend the concept of 'Trustworthy AI' (e.g. Weydner-Volkmann & Feiten 2021; Bleher & Hummel 2021). More prominently, the concept of 'Trustworthy AI' has already become ubiquitous, and we have passed the stage of principles and guidelines, now expecting regulations and legislation. The larger AI communities will probably continue to use this term.

Self-reflective AI ethicists understand that in the context of technologies, the field of ethics "has become increasingly institutionalized, instrumentalized, and professionalized" (Green 2022). Additionally, the field of ethics of AI lacks any real-life mechanisms to support its own normative claims (Hagendorff 2020: 99). As much as ethicists know that the objective of ethics, unlike laws and regulations, is not to impose norms or to ensure compliance (Rességuier & Rodrigues 2020), I believe scholars in this field do wish to make an impact beyond the mere terminology used internally.

4.2. Scholarly Activism

If the criticisms expressed by trust scholars were to be summarized into a practical suggestion, it would be ensuring that individuals and institutions that regard AIs are trustworthy. Beyond reducing trust in AI to trust in the individuals and institutions, trust scholars can focus on the democratic and political culture surrounding AIs as trustworthy (Rieder, Simon, & Wong 2021). For example, they can emphasize the efforts of striving for further public engagement and deliberation about the development and funding of AIs (Vesnic-Alujevic, Nascimento, & Polvora 2020).

Additionally, trust scholars can point to occurrences of distrust. Distrust, as Braun, Bleher & Hummel (2021) argue, is not merely disappointment or broken trust. Instead, distrust plays an important and constructive social role: it opens and advances societal negotiations and the ability to problematize and recognize critical social aspects. Distrust indicates shifts in power and advances societal negotiations.

AI ethicists and trust scholars can uncover and criticize structural injustice, such as pointing to the concentration of power in the AI market, the usage of AI by exploitative corporations, or highlight issues which regard monitoring, profiling, tracking, measuring, and predicting. When it comes to democratic and political culture, distrust is a complementary phenomenon to trust.

AI ethicists and trust scholars are able to stress the importance of activism in forming trust between citizens and those who largely influence the development, implementation, maintenance, and auditing of AIs. One way to do so is to amplify the voices of communities that were systematically excluded from shaping AIs, such as women, people who are disabled, LGBTQ+, Indigenous, Black, or poor (Kalluri 2020). Another way is to engage with social initiatives. One example is *ForHumanity*[6] – an all-volunteer non-profit organization advocating that "everyone animating the debate on AI systems is necessary for building a credible infrastructure of trust" and runs high-impact projects. By engaging with social justice issues, being action-oriented, and challenging the status quo, the difference between scholarly inquiry and activism mergers into 'scholar activism' (Ramasubramanian & Sousa 2021).

5. Conclusion

[6] ForHumanity's website: https://forhumanity.center.

This paper explained why trust scholars and AI ethicists criticism the concept of 'Trustworthy AI'. In section 1, I distinguished between criticisms of HLEG's 'Trustworthy AI' initiative and criticisms of the concept 'Trustworthy AI'. In section 2, I presented the theoretical roots of opposing the term ‘Trustworthy AI’. I pointed to the anthropocentric view of trust within social epistemology, which argues that only human agents can be objects of trust, and therefore trust in AI is reduced to issues of trust in the institutions and people behind the technology. I traced the origins of this view to three ideas underlying the distinction between genuine trust and mere reliance.

In section 3, four main criticisms about the concept of 'Trustworthy AI' were reviewed. The criticisms included Bryson's (2018) dual argument about peers and accountability; Margit Sutrop's (2019) argument about social cognition; Ryan's (2020) argument about goodwill and the ability to normatively commit; and Rieder, Simon, and Wong (2020) analysis of the democratic and political aspects of the concept. The section ended by detailing the consequences of associating AI systems with human-like or institution-like characteristics – such as 'trustworthiness'. I argued that misplacing trust - i.e. directing trust at the wrong objects, leads to wrong attribution of moral and legal responsibilities.

Lastly, in section 4, the question of "what's next" was raised. First, I suggested that the field of ethics of AI grew too big to change its terminology from the paradigm of 'Trustworthy AI' to 'Reliable AI'. I then proposed that if we wish to be practical, those working on the topic of trust and ethics of AI can focus on social issues of trust. Since distrust is also essential to uncovering power relations, it is worth focusing on this concept, too.

Ultimately, trust and AI ethics is about power, social justice, and activism. In many cases, dominant institutions and those who impact AIs influence power dynamics while assuming or asking for our trust. In this respect, the concepts of ‘trust’ and ‘distrust’ in democratic institutions are about power. If community-driven and social justice-oriented ethicists of AI and trust scholars wish to stay relevant, they must become scholarly activists, who uncover power relations, challenge unfair status quo, and suggest democratic ways to form trust relations.

Statements and Declarations

No potential conflict of interest was reported by the author(s).

References

AlgorithmWatch. (2019). No red lines: Industry defuses ethics guidelines for artificial intelligence. https://algorithmwatch.org/en/industry-defuses-ethics-guidelines-for-artificial-intelligence/

Article 19. (2019). Governance with teeth: How human rights can strengthen FAT and ethics initiatives on artificial intelligence. April 17, 2019. https://www.article19.org/resources/governance-with-teeth-how-human-rights-can-strengthen-fat-and-ethics-initiatives-on-artificial-intelligence/

Baier, A. (1986). Trust and antitrust. *Ethics,* 96(2), 231-260. https://doi.org/10.1086/292745

Braun, M., Bleher, H., & Hummel, P. (2021). A leap of faith: is there a formula for "Trustworthy" AI?. *Hastings Center Report*, 51(3), 17-22. https://doi.org/10.1002/hast.1207

Bryson, J. J. (2017). The meaning of the EPSRC principles of robotics. *Connection Science*, 29(2), 130-136. https://doi.org/10.1080/09540091.2017.1313817

Bryson, J. J. (2018). "AI & Global Governance: No One Should Trust AI," November 13, 2018, United Nations University, Centre for Policy Research, https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html

CAICT [China Academy of Information and Communications Technology]. (2021). White Paper on Trustworthy Artificial Intelligence. www.caict.ac.cn/english/research/whitepapers/202110/t20211014_391097.html

Coeckelbergh, M. (2012). "Can we trust robots?", *Ethics and Information Technology,* 14(1), 53-60. https://doi.org/10.1007/s10676-011-9279-1

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, 26(4), 2051-2068. https://doi.org/10.1007/s11948-019-00146-8

Danaher, J. (2020). Welcoming robots into the moral circle: a defence of ethical behaviourism. Science and Engineering Ethics, 26(4), 2023-2049. https://doi.org/10.1007/s11948-019-00119-x

Davies, Jamie. (2019). Europe publishes stance on AI ethics, but don't expect much, telecoms.com news 28 June 2019. https://telecoms.com/498190/europe-publishes-stance-on-ai-ethics-but-dont-expect-much

De Filippi, P., Mannan, M., & Reijers, W. (2020). "Blockchain as a confidence machine: The problem of trust & challenges of governance", Technology in Society 62. https://doi.org/10.1016/j.techsoc.2020.101284

Dotan, R. (2021). The Proliferation of AI Ethics Principles: What's Next?, MAIEI. https://montrealethics.ai/the-proliferation-of-ai-ethics-principles-whats-next/

Dubber, M. D., Pasquale, F., & Das, S. (Eds.). (2020). *The Oxford Handbook of Ethics of AI*. Oxford Handbooks.

EC [European Council]. (2021). Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts. (Document 52021pc0206). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Farina, L. (2022). Sven Nyholm, Humans and Robots; Ethics, Agency and Anthropomorphism. *Journal of Moral Philosophy*, 19(2), 221-224. https://doi.org/10.1163/17455243-19020007

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

Floridi, L (2019). Translating principles into practices of digital ethics: Five risks of being unethical. Philosophy & Technology, 32: 185–193. https://doi.org/10.1007/s13347-019-00354-x

Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. In Ethics, Governance, and Policies in Artificial Intelligence (pp. 5-17). Springer, Cham. https://doi.org/10.1007/978-3-030-81907-1_2

Freiman, O. (2014). Towards the Epistemology of the Internet of Things: Techno-Epistemology and Ethical Considerations Through the Prism of Trust, *International Review of Information Ethics* 22(2): 6-22. https://doi.org/10.29173/irie115

Freiman, O. (2021). The Role of Knowledge in the Formation of Trust in Technologies. Ph.D. Dissertation, Bar-Ilan University.

Freiman, O., & Miller, B. (2020). "Can Artificial Entities Assert?",In: S. Goldberg (ed.), The Oxford Handbook of Assertion. Oxford University Press. https://academic.oup.com/edited-volume/34275/chapter-abstract/290604123

Gießler, S., Spielkamp, M, Ferrario, A., Christen, M., Shaw, D., Schneble, C. (2019). 'Trustworthy AI' is not an appropriate framework. Algorithm Watch. https://algorithmwatch.org/en/trustworthy-ai-is-not-an-appropriate-framework/

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627-660. https://doi.org/10.5465/annals.2018.0057

Green, B. (2022). The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice. The Digital Humanist, February 25, 2022. https://thedigitalhumanist.org/the-contestation-of-tech-ethics-a-sociotechnical-approach-to-technology-ethics-in-practice

Gunkel, D. J. (2018a). The other question: can and should robots have rights?. Ethics and Information Technology, 20(2), 87-99. https://doi.org/10.1007/s10676-017-9442-4

Gunkel, D. J. (2018b). *Robot Rights*. MIT Press.

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

Hardin, R. (1993). The street-level epistemology of trust. Politics & Society 21(4): 505-529. https://doi.org/10.1177/0032329293021004006

Hatherley, J. J. (2020). Limits of trust in medical AI. Journal of Medical Ethics, 46(7), 478-481. https://doi.org/10.1136/medethics-2019-105935

Hawley, K. (2014). Trust, distrust and commitment. Noûs, 48(1), 1-20. https://doi.org/10.1111/nous.12000

HLEG. (2018). Draft Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/draft-ethics-guidelines-trustworthy-ai

HLEG. (2019). Ethics guidelines for trustworthy AI . https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Humphreys, P. (2009). Network Epistemology. Episteme 6(2): 221-229. https://doi.org/10.3366/e1742360009000653

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

Jones, K. (1996). Trust as an affective attitude. *Ethics*, 107(1), 4-25. https://doi.org/10.1086/233694

Jones, K. (2012). Trustworthiness. *Ethics*, 123(1), 61-85. https://doi.org/10.1086/667838

Kalluri, P. (2020). Don't Ask if Artificial Intelligence Is Good or Fair, Ask How It Shifts Power.' *Nature* 583. https://doi.org/10.1038/d41586-020-02003-2

Keymolen, E. (2016). Trust on the line: a philosophical exploration of trust in the networked era. Ph.D. Dissertation, Erasmus University Rotterdam.

Kontogiorgos, D., et al. (2019). "The effects of anthropomorphism and non-verbal social behaviour in virtual assistants", in: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, pp. 133-140. https://doi.org/10.1145/3308532.3329466

Metzinger, T. (2019). Ethics Washing Made in Europe (Der Tagesspiegel, 2019), https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

Metzinger, T., & Coeckelbergh, M. (2020). Europe needs more guts when it comes to AI ethics. Tagesspiegel BACKGROUND, 16. April 2020. https://background.tagesspiegel.de/digitalisierung/europe-needs-more-guts-when-it-comes-to-ai-ethics

Miller, B., & Freiman, O. (2020). "Trust and Distributed Epistemic Labor", In: J. Simon (ed.), *The Routledge Handbook on Trust and Philosophy*. Routledge.

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

NAII [National Artificial Intelligence Initiative]. (2021). Advancing Trustworthy AI. https://www.ai.gov/strategic-pillars/advancing-trustworthy-ai/

Nguyen, T. C. (Forthcoming). "Trust as an Unquestioning Attitude", In: Oxford Studies in Epistemology.

Nickel, P. J. (2013). "Trust in Technological Systems", in: De Vries, M. J., Hansson, S. O., and Meijers, A. W. (eds.), Norms in technology, pp. 223-237. Springer. https://doi.org/10.1007/978-94-007-5243-6_14

Nickel, P. J. (2022). Trust in medical artificial intelligence: a discretionary account. *Ethics and Information Technology*, 24(1), 1-10. https://doi.org/10.1007/s10676-022-09630-5

Nickel, P.J., Franssen, M., and Kroes, P. (2010). Can We Make Sense of the Notion of Trustworthy Technology? Knowledge, Technology & Policy, 23(3-4), pp. 429–444. https://doi.org/10.1007/s12130-010-9124-6

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism.* Rowman & Littlefield Publishers.

Origgi, G. (2008). Qu'est-ce que la confiance? Paris: VRIN.

Peukert, C., & Kloker, S. (2020). Trustworthy AI: How Ethicswashing Undermines Consumer Trust. WI2020 Zentrale Tracks, 1100–1115. https://doi.org/10.30844/wi_2020_j11-peukert

Pitt, J. C. (2010). It's not about technology. *Knowledge, Technology and Policy* 23(3-4):445-454. https://doi.org/10.1007/s12130-010-9125-5

Ramasubramanian, S., & Sousa, A. N. (2021). Communication scholar-activism: conceptualizing key dimensions and practices based on interviews with scholar-activists. Journal of Applied Communication Research, 49(5), 477-496. https://doi.org/10.1080/00909882.2021.1964573

Renda, A. (2020). Europe: toward a policy framework for trustworthy AI. The Oxford Handbook of Ethics of AI, 649-666. https://doi.org/10.1093/oxfordhb/9780190067397.013.41

Rieder, G., Simon, J., & Wong, P. H. (2021). Mapping the stony road toward trustworthy AI: expectations, problems, conundrums. Machines We Trust: Perspectives on Dependable AI. Cambridge, MA: MIT Press, 2021. https://doi.org/10.7551/mitpress/12186.003.0007

Ryan, M. (2020). In AI we trust: ethics, artificial intelligence, and reliability. Science and Engineering Ethics, 26(5), 2749-2767. https://doi.org/10.1007/s11948-020-00228-y

Schiff, D., J. Borenstein, J. Biddle, and K. Laas (2021). AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection. *IEEE Transactions on Technology and Society* 2(1): 31-42. DOI: http://dx.doi.org/10.1109/TTS.2021.3052127

Simon, J. (2010). The Entanglement of Trust and Knowledge on the Web. *Ethics and Information Technology,* 12(4), 343-355. https://doi.org/10.1007/s10676-010-9243-5

Simon, J. (2013). “Trust”, in: D. Pritchard (ed.). Oxford Bibliographies in Philosophy. Oxford University Press. https://doi.org/10.1093/obo/9780195396577-0157

Stamboliev, E. (2020). Robot Rights by David J. Gunkel. Leonardo, 53(1), 110-111. https://doi.org/10.1162/leon_r_01849

Sutrop, M. (2019). Should we trust artificial intelligence?. Trames, 23(4), 499-522. https://doi.org/10.3176/tr.2019.4.07

Tallant, J. (2019). You can trust the ladder, but you shouldn't. Theoria, 85(2), 102-118. https://doi.org/10.1111/theo.12177

Tamir, P., & Zohar, A. (1991). “Anthropomorphism and teleology in reasoning about biological phenomena”, Science Education 75(1): 57-67. https://doi.org/10.1002/sce.3730750106

Tavani, H. T. (2018). Can social robots qualify for moral consideration? Reframing the question about robot rights. Information, 9(4), 73. https://doi.org/10.3390/info9040073

Vesnic-Alujevic, L., Nascimento, S., & Polvora, A. (2020). Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks. Telecommunications Policy, 44(6), 101961. https://doi.org/10.1016/j.telpol.2020.101961

Weydner-Volkmann, S., & Feiten, L. (2021). Trust in technology: interlocking trust concepts for privacy respecting video surveillance. *Journal of Information, Communication and Ethics in Society*, 19(4), 506-520. https://doi.org/10.1108/jices-12-2020-0128

Wilholt, T. (2009) “Bias and values in scientific research,” Studies in History and Philosophy of Science 40(1): 92-101. https://doi.org/10.1016/j.shpsa.2008.12.005

Winfield, A. F., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180085. https://doi.org/10.1098/rsta.2018.0085