Draft version © Rieder, G., Simon, J. and Wong, P.-H., to be published in: Pelillo, M. and Scantamburlo, T. (Eds.). *Machines We Trust: Perspectives on Dependable AI*. Cambridge, MA: MIT Press, forthcoming 2021.

Mapping the Stony Road toward Trustworthy AI: Expectations, Problems, Conundrums

Gernot Rieder I Judith Simon I Pak-Hang Wong

Abstract

The notion of trustworthy AI has been proposed in response to mounting public criticism of AI systems, in particular with regard to the proliferation of such systems into ever more sensitive areas of human life without proper checks and balances. In Europe, the High-Level Expert Group on Artificial Intelligence has recently presented its *Ethics Guidelines for Trustworthy AI*. To some, the guidelines are an important step for the governance of AI. To others, the guidelines distract effort from genuine AI regulation. In this chapter, we engage in a critical discussion of the concept of trustworthy AI by probing the concept both on theoretical and practical grounds, assessing its substance and the feasibility of its intent. We offer a concise overview of the guidelines and their vision for trustworthy AI and examine the conceptual underpinnings of trustworthy AI by considering how notions of 'trust' and 'trustworthiness' have been discussed in the philosophical literature. We then discuss several epistemic obstacles and moral requirements when striving to achieve trustworthy AI in practice before concluding with an argument in support of the establishment of a *trustworthy AI culture* that respects and protects foundational values.

Keywords

Trust, Trustworthy AI, Fairness, Transparency, Accountability, AI Ethics

1. Introduction

Recently, the notion of trustworthy AI has become a key term in debates about the potential ethical implications of the widespread adoption of artificial intelligence (AI) systems in different societal domains. In Europe, the High-Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission (EC) has presented its *Ethics Guidelines for Trustworthy AI* (AI HLEG, 2019), literally placing the notion front and center. Shortly after, the Organisation for Economic Co-operation and Development published a *Recommendation on AI* (OECD, 2019) that emphasizes the need for an "international co-operation for trustworthy AI". In the United States, the White House Office of Science and Technology Policy has pointed to the importance of developing approaches for the trustworthy creation and adoption of new AI technologies (OSTP, 2019), and the Chinese *Beijing AI Principles* (BAAI, 2019) hold that AI research and development should adopt ethical design approaches to make systems trustworthy.

These calls for trustworthy AI come at a time when the critique of automated decision-making has become more public and pronounced (see, e.g., Eubanks, 2018; O'Neil, 2016). In particular examples of algorithmic bias and discrimination in criminal sentencing (Angwin et al., 2016), recruitment (Dastin, 2018) and facial analysis (Buolamwini and Gebru, 2018) have garnered public attention.[1] What worries many is that the proliferation of such systems into ever more sensitive areas of human life appears to progress without proper checks and balances. As an extensive report on the use of decision-making software in the EU for instance notes, "it is doubtful that many oversight bodies in place have the expertise to analyze and probe modern automated decision-making systems and their underlying models for risk of bias, undue discrimination, and the like." (Spielkamp, 2019: 15)

[1] For more on the issue of bias in machine learning, see chapter 4 by Mireille Hildebrandt in this volume.

Similar concerns have been raised in the U.S., with commentators noting a lack of tech literacy amongst policymakers, making it difficult for authorities to perform their regulatory and oversight function (see, e.g., Lamberth, 2019).

Against this background, attempts to govern the rise of AI via soft law approaches rather than hard law measures may seem like the most viable course of action. After all, how to regulate a technology that has many different applications, evolves quickly and involves problems that are not always easy to solve? The fact that AI research and development is now perceived as a major area of competition between the United States, China and Europe factors in as well, as heavy-handed regulation is often seen as a threat to technological advancement and industrial growth, potentially undermining a nation's ability to successfully compete in global markets (see, e.g., Castro et al., 2019). Thus, at least for the time being, lighter-touch approaches continue to be the governance strategy of choice, with ethical guidelines and non-binding codes of conduct – rather than legally enforceable rules – taking center stage (see Marchant, 2019).

Reactions to the focus on soft law initiatives have been mixed, in particular with regard to the EC-commissioned *Ethics Guidelines for Trustworthy AI* (AI HLEG, 2019). Whereas some have greeted the guidelines as a "system of checks and balances […] to prevent abuses of power" (Loritz, 2019) or a "good starting point" for figuring out "whether new AI systems are ethical" (Samuel, 2019), others have lamented that "the guidelines do not seem to have any substance to them" and that "AI regulation needs action not philosophical thinking" (Davies, 2019). In a controversial opinion piece, an academic member of the expert group that worked on the guidelines for nine months called the result "a compromise" around an industry-invented "marketing narrative" that uses "ethics debates as elegant public decorations for a large-scale investment strategy" (Metzinger, 2019). According to the author, what can be observed is a form of "ethics washing" where industry "organizes and cultivates ethical debates to buy time – to distract the public and to prevent or at least delay effective regulation and policy-making." (ibid.) Others have made similar arguments, adding to the guidelines' controversial reception (see, e.g., Hidvegi and Leufer, 2019).

In this chapter, we engage in a critical discussion of the concept of trustworthy AI as it is outlined in the High-Level Expert Group's ethics guidelines (AI HLEG, 2019). However, rather than questioning the conditions of its development – that is, the composition and the timing of the expert group – we shall probe the concept both on theoretical and practical grounds, assessing its substance and the feasibility of its intent. More specifically, after a concise overview of the guidelines and their vision for trustworthy AI (section 2), we take a closer look at the conceptual underpinnings of trustworthy AI by considering how notions of 'trust' and 'trustworthiness' have been discussed in the philosophical literature (section 3.1). We then identify several obstacles when striving to achieve trustworthy AI in practice, debating practitioners' limits of competence when balancing trade-offs between competing values (section 3.2) before outlining moral obstacles that may render the achievement of trustworthy AI difficult (section 3.3). We conclude by arguing for the establishment of a *trustworthy AI culture* that respects and protects foundational values (section 4).

2. The High-Level Expert Group's *Ethics Guidelines for Trustworthy AI*

The EC-commissioned *Ethics Guidelines for Trustworthy AI* (AI HLEG, 2019) were developed by a 52-member expert group comprising representatives from academia, civil society and industry. [2]

[2] For more information on the High-Level Expert Group and its exact composition, see the European Commission website at https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.

Assembled in June 2018, the group released a first draft of the guidelines in December 2018 and, after an open consultation process that generated feedback from more than 500 contributors,[3] presented a revised version in April 2019. The European Commission has since welcomed the work of the expert group and considers the ethics guidelines a "valuable input for its policy-making" (EC, 2019: 4). More specifically, the guidelines are meant to inform and support the European AI strategy, which holds that an "environment of trust and accountability around the development and use of AI is needed" (EC, 2018: 13). Consequently, in order to facilitate public acceptance and increase Europe's competitive advantage, the "trustworthiness of AI should be ensured", which implies that "AI applications should not only be consistent with the law, but also adhere to ethical principles […,] empower citizens and respect their fundamental rights" (EC, 2019: 2). And this is where the *Ethics Guidelines for Trustworthy AI* come into play.

According to the guidelines,[4] trustworthy AI has three components: it should be *lawful*, complying with applicable laws and regulations; it should be *ethical*, ensuring adherence to ethical principles and values; and it should be technically and socially *robust*, performing in a safe, secure and reliable manner. While all three components are regarded as necessary for achieving trustworthy AI, the guidelines do not explicitly deal with the first component (lawful AI), but instead strive to offer guidance on how to foster and secure the second and third component (ethical and robust AI) by putting forward a set of *principles* and *requirements* that should be respected in the development, deployment and use of AI systems. Concretely, the guidelines specify four ethical principles and seven key requirements that should be met throughout a system's life cycle.

The four ethical principles are: (1) *respect for human autonomy*, meaning that humans interacting with AI systems must be able to keep full and effective self-determination over themselves without being unjustifiably subordinated, deceived or conditioned; (2) *prevention of harm*, which implies that AI systems should neither cause nor exacerbate harm, protecting human dignity with particular attention to vulnerable persons and power asymmetries; (3) *fairness*, meaning that AI systems must ensure that individuals and groups are not subject to unfair bias, discrimination or stigmatization (substantive dimension) and able to contest and seek effective redress against decisions made by such systems (procedural dimension); (4) *explicability*, meaning that processes and their purpose need to be transparent and that decisions must be explainable to those directly and indirectly affected, especially if the potential consequences of the decision are severe (sectorial approach).

These four principles, conceptually based on rights and values laid down in legal documents such as the Treaty on European Union[5] and the Charter of Fundamental Rights of the EU[6], are meant to serve as ethical imperatives that AI practitioners should always strive to adhere to. As such, they aim beyond mere legal compliance, seeking to provide guidance in identifying what *should* be done rather than what legally *can* be done with (AI) technology. The guidelines, however, acknowledge that despite offering guidance, the principles remain abstract ethical prescriptions that, in order to support

[3] For more information on the consultation process, see the European Commission website at https://ec.europa.eu/digital-single-market/en/news/over-500-comments-received-draft-ethical-guidelines-trustworthy-artificial-intelligence.

[4] The following paragraphs provide a concise overview of the framework for trustworthy AI as outlined in the High-Level Expert Group's ethics guidelines. The goal is to familiarize readers with the framework's main elements before delving into a deeper discussion in the ensuing sections. For full details of the framework, please consult the original guidelines document.

[5] Consolidated Version of the Treaty on European Union (2012) *Official Journal of the European Union*, C 326, 26 Oct., pp. 13-45.

[6] Charter of Fundamental Rights of the European Union (2012) *Official Journal of the European Union,* C 326, 26 Oct, pp. 391-407.

the actual realization of trustworthy AI, must be translated into concrete requirements that developers should implement into their designs, deployers should ensure to meet in their products and services, and end-users and the society at large should be aware of so as to be able to request that they are fulfilled.

In addition to the ethical principles the guidelines specify seven key requirements that must also be met: (1) *human agency and oversight*, allowing users to make informed decisions regarding AI systems and creating effective mechanisms for monitoring and human intervention; (2) *technical robustness and safety*, which requires that AI systems are developed with a preventative approach to risks, behaving reliably as intended while minimizing unintentional and preventing unacceptable harm; (3) *privacy and data governance*, ensuring the protection and lawful processing of personal data while assessing the quality of datasets and implementing access restrictions; (4) *transparency*, including the proper documentation of datasets and analytical procedures, the explainability of both a system's decision-making process and the rationale for deploying it, the identifiability of a system during interaction, and the communication of a system's capacities and limitations appropriate to the use case at hand; (5) *diversity, non-discrimination and fairness*, aiming for the avoidance of unfair bias through oversight processes, greater involvement of affected stakeholders, more inclusive design processes, and the application of universal design and accessibility standards; (6) *societal and environmental well-being*, taking into account how systems affect institutions, democracy and society at large and fostering research into sustainable and environmentally friendly AI; (7) *accountability*, enabling evaluation by internal or external auditors, facilitating the identification of potential negative consequences through impact assessments during all stages of a system's life cycle, willingness to accept trade-offs between relevant interests and potential risks to the point where the development, deployment and use of a system cannot proceed, and mechanisms for redress when unjust adverse impacts have occurred.

The guidelines also specify a number of technical and non-technical methods of how these requirements can be implemented. The technical methods include (1) the development of *architectures* and *procedures* that "white list" rules that a system should follow and "black list" behaviors that a system must not exhibit, (2) the application of *values-, privacy-* and *security-by-design* approaches that ask companies to consider the impact of their systems from the very start, (3) research on *explanation methods* that provide insight into why a system behaved in a certain way and provided a given interpretation, (4) the incorporation of *testing* and *validation processes* that include all components of an AI system and ensure that the system behaves as intended throughout its entire life cycle, and (5) the definition of appropriate *quality of service indicators* to ensure that systems have been tested and developed with security and safety considerations in mind. The non-technical methods include (1) *regulatory frameworks*, in acknowledgment of the need to revise and adapt existing or introduce new legislation, (2) *codes of conduct*, with organizations documenting their intentions, underwritten by desirable values such as fundamental rights, transparency or the avoidance of harm, (3) *standards* that offer the ability to recognize and encourage ethical conduct in AI designs, with a possible "Trustworthy AI" label confirming by reference to specific technical standards that a system, for instance, adheres to safety, technical robustness and transparency, (4) *certifications* for organizations that can attest to a broader public that an AI system is transparent, accountable and fair without replacing corporate responsibility, (5) internal and external *governance frameworks* – e.g., a person, panel or board – set up by organizations to complement legal oversight, (6) *communication*, *education* and *training* to foster basic AI literacy and an ethical mind-set across society, (7) *stakeholder participations* and *social dialogue* to discuss the use and impact of AI systems and data analytics, and (8) increasing *diversity* among the teams that design and develop, test and maintain, deploy and procure AI systems, not only in terms of gender, culture or age, but also in terms of professional background and skill sets.

The guidelines conclude with an extensive assessment list for the operationalization of trustworthy AI, geared specifically toward developers and deployers of AI systems. The list contains a set of questions based on the seven key requirements outlined above and is meant to encourage reflection on how trustworthy AI be achieved and the steps that should be taken in this regard. Lastly, the guidelines emphasize that trustworthy AI is not about ticking boxes, but about continuously identifying requirements, evaluating solutions and ensuring improved outcomes throughout an AI system's life cycle.

3. Discussion

As mentioned in the introduction, the High-Level Expert Group's guidelines have been welcomed by the European Commission that seeks to ensure that "this guidance can be tested and implemented in practice" (EC, 2019: 7). To this end, the Commission has invited stakeholders who develop or use AI to provide feedback on the assessment list to evaluate the guidelines' workability and feasibility.[7] The Commission sees this push for ethical AI as a win-win proposition: While "guaranteeing the respect for fundamental values and rights is not only essential in itself", it also "facilitates acceptance by the public and increases the competitive advantage of European AI companies by establishing a brand of human-centric, trustworthy AI known for ethical and secure products." (ibid.: 8) This implies that, ultimately, the Commission's push for trustworthy AI is about more than ensuring that AI R&D is conducted in due consideration of the fundamental rights and values of the European Union. It is also about the strategic positioning of a "brand" that can be promoted – and sold – in European and global markets. But this focus on the commodification of trustworthy AI is where things become difficult, both on theoretical and practical grounds. In this section, we shall first examine the conceptual underpinnings of trustworthy AI by considering how notions of 'trust' and 'trustworthiness' have been discussed in the philosophical literature before continuing to identify a number of epistemic and moral obstacles when striving to achieve trustworthy AI in practice.

3.1 On 'Trustworthy AI' and 'Trusting AI'

As the objective of trustworthy AI is to encourage people's trust in AI systems, how trustworthy AI will be realized depends crucially on how 'trust' and 'trustworthiness' are interpreted. So, it is essential for our discussion to first clarify the conceptual underpinnings of trustworthy AI. The concepts of 'trust' and 'trustworthiness' have been extensively discussed by philosophers (see, e.g., Simon, 2013). In philosophical analyses of trust, 'trust' is ordinarily formulated as a three-place relation involving a trustor *A*, a trustee *B*, and either a domain of interaction or a specific good *P*. 'Trustworthiness' or 'being trustworthy' is then defined by the character of being merit of trust. Philosophers often distinguish between (mere) reliance and (genuine) trust, and relatedly between reliability and trustworthiness (see, e.g., Baier, 1986). While reliance is defined in terms of the *rational* expectation of a dependent person about the person (or entity) being depended upon, 'trust' is often defined *not* only in terms of *rational* expectation but in *moral* terms. In her seminal article "Trust and Antitrust", Baier (1986: 235) defines trust as "accepted vulnerability to another's possible but not expected ill will

[7] The Assessment List of the Ethics Guidelines for Trustworthy AI is available at https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60440.

(or lack of good will) towards one". A violation of trust induces in the trustor a feeling of betrayal and not merely one of disappointment. Thus, if trustworthy AI aims to encourage people's *trust* in AI systems but not mere *reliance* on them, the concepts of 'trust' and 'trustworthiness' need to be interpreted *morally*.

In their discussion on the possibility of *trust in technology*, Nickel et al. (2010: 431-433) helpfully identify two families of accounts of trust, namely (a) pure rational-choice accounts, in which 'trust' is conceptualized as a rational, cost-benefit calculation regarding the effectiveness of relying on another person (or entity) in completion of a specific task, and (b) "motivation-attributing" accounts, in which 'trust' requires the trustor to attribute some motivations to the trustee). They rightly point out that pure rational-choice accounts *cannot* distinguish between reliance and trust (and, in the same vein, between reliability and trustworthiness) because in these accounts trust is reduced to rational expectation of the performance of action by the person or entity being depended upon, and thus it lacks the *moral* dimension required by genuine trust. Accordingly, the concept of 'trust' in trustworthy AI cannot be based on a pure rational-choice account if it intends to distinguish between *(merely) reliable* AI and *trustworthy* AI. Indeed, the guidelines seem to acknowledge the insufficiency of a pure rational-choice account, as ensuring the proper- and well-functioning of AI systems is only one of the seven requirements the guidelines specify, i.e. *(2) technical robustness and safety*, which requires that AI systems "includ[e] resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility" (EC, 2019: 14).

It thus leaves us with "motivation-attributing" accounts of trust as the conceptual basis of trustworthy AI, but these accounts are not without problem when applied to trustworthy AI. Particularly, since "motivation-attributing" accounts require the trustor to attribute *some* motivations to the trustee with regards to their values and interests, non-human entities like AI systems are not appropriate targets for attribution of motivations as they do not have the right type of capacities to morally account for the trustor's values and interests (Nickel *et al.*, 2010: 435; also see Jones, 1996; cf. Tallant, 2019). Insofar as current AI systems cannot exhibit such *mental* and *moral* capacities to respond to the trustor's values and interests, trustworthy AI also cannot be based on a "motivation-attributing" account.[8]

There are, however, potential ways to reframe the discussion of trustworthy AI to render the "motivation attributing" account *plausible* in the discussion. For example, Nickel *et al.* (2010) explore the plausibility of people *trusting* technology in a *derived sense*. They note that technologies, especially complex technical objects, are situated and implemented in a network that includes both technical object(s) and *human agents* who design, manufacture, manage and/or operate the technical object(s) in question. In accordance with the "motivation-attributing" account, it is *plausible* for people to *trust* the technical objects indirectly, i.e. to *trust* socio-technical systems through trusting the human agents who design, manufacture, manage or operate them and who are capable of accounting for the trustors' values and interests. Accordingly, *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 designing, creating, managing and/or operating them (see, e.g., Ananny 2016). In effect, this broad understanding of AI systems seems to be adopted in the guidelines, as it is stated that "while 'Trust' is usually not a property ascribed to machines, this document aims to stress the importance of being able to trust not only in the fact that AI systems are legally compliant, ethically

[8] It remains an open question whether AI systems can be moral *agents* that have mental and moral capacities. For an overview of the qualities of AI systems to obtain moral agency, see Gunkel (2018).

adherent and robust, but also that such trust can be ascribed to all people and processes involved in the AI system's life cycle" (EC, 2019: 38).

The discussion so far has been to establish a *plausible* interpretation of trust for trustworthy AI. Our claim is that trust cannot be based on a pure rational-choice account, in which trust and trustworthiness are reduced to (mere) reliance and reliability, and that trustworthy AI should refer to the "motivation-attributing" account. Doing so, however, requires us to take the *network* of AI-based technologies and human agents, who are in various ways involved with the AI systems, as the unit of analysis. Yet, even if trust in AI systems is *plausible*, it remains an open question as to whether and when people's trust in AI systems is justified or valuable. As O'Neill rightly argues that "[t]rust is valuable only when directed to agents and activities that are trustworthy" (2018: 293). In order to answer this question, it is essential to specify the criteria for AI systems, understood as socio-technical systems, to be trustworthy.

We have stated earlier that 'trustworthiness' or 'being trustworthy' is defined by the character of being merit of trust. There are two types of criteria shared by many accounts of trust and trustworthiness as a moral notion, namely *competence* and some kind of *trust-responsiveness* (Ruokonen, 2013). To be competent, the trustee needs to be capable of fulfilling the trustor's expectation with respect to the tasks she is entrusted. To be trust-responsive, on the other hand, the trustee needs to in some ways account for the trustor's values and interests as she completes the tasks. For example, Baier (1986) refers to the trustee's good will (or lack of ill will) towards trustors, whereas Jones defines trustworthiness as follows: "*B* is trustworthy with respect to *A* in domain of interaction *D*, if and only if she is *competent* with respect to that domain, and she *would* take the fact that A is counting on her, were A to do so in this domain, to be a *compelling* reason for acting as counted on" (Jones 2012: 70-71; original emphasis).

Purely game-theoretic accounts of trust have also been challenged within philosophy of science. In his influential article *The Role of Trust for Knowledge*, Hardwig (1991) argues that in order to know, in particular but not only in science, we need to trust others in regards to their truthfulness, their competence and their adequate epistemic self-assessment, i.e. the second-order competence to also know the limits of their competence.[9] Accordingly, trustworthiness is considered to have a *moral component* (being truthful) and *an epistemic* component (being *knowledgeable* and *free from epistemic vices*, such as the tendency to misjudge or deceive oneself about her own capacities) (Hardwig, 1994: 88-89; cf. O'Neill, 2018).

To summarize, there are several epistemic and moral requirements which human agents must fulfill to be considered trustworthy: on the epistemic side, they must be competent and know the limits of their competence; on the moral side, they must be trust-responsive and truthful. What do these requirements mean for AI systems?

3.2 The Epistemic Obstacles for Trustworthy AI

As outlined, the ethics guidelines specify a number of principles and requirements that practitioners should always strive to adhere to in order to ensure that AI systems are developed, deployed and used in a trustworthy manner. In other words, the ethics guidelines have created an expectation of the required level of competence for AI systems to be trustworthy. As argued in the previous section, trustworthiness implies that the trustee must have sufficient *competence* to effectively fulfill the

[9] It should be noted that being competent requires the trustee to have relevant skills, know-how, experience, and propositional knowledge.

trustor's expectations. With respect to the guidelines, this means that in order for a socio-technical system involving AI applications to be trustworthy, the developing entity – i.e., a company, organization or institution – must in fact be able to meet the stated requirements. So, trustworthy AI may be difficult to achieve if the requirements in the guidelines are difficult to satisfy. In effect, the guidelines are well aware that satisfying the requirements is often not an easy or straightforward process, pointing to "tensions" that can arise between the ethical principles and "trade-offs" that will have to be considered when trying to implement the requirements.

For instance, tensions may arise between efforts to reduce crime through the use of AI systems for predictive policing (related to the *principle of prevention of harm*) and ethical imperatives to safeguard individual rights and freedoms (most closely linked to the *principle of respect for human autonomy*). Likewise, trade-offs might have to be made between an AI system's decision being fully explainable (part of the guidelines' *requirement of transparency*) and the wish for judgments to be as accurate as possible (part of the *requirement of technical robustness and safety*) as the methods that produce the best results may employ machine learning processes that cannot be easily grasped by human reasoning. The guidelines advise that AI practitioners should "approach ethical dilemmas and trade-offs via reasoned, evidence-based reflection" (AI HLEG, 2019: 13), addressing them in a "rational and methodological manner within the state of the art." (ibid., 13) This seems prudent, especially since the guidelines concede that "[i]n situations in which no ethically acceptable trade-offs can be identified, the development, deployment and use of the AI system should not proceed in that form" (ibid.). Yet, it does beg the question of whether the level of authority conferred to AI practitioners does not overstate their competence in making difficult value-based decisions. Of course, the point here is not to deny developers' knowledgeability beyond the purely technical, but to emphasize that when working toward more trustworthy AI environments, awareness of the potential limits of competence of *any* particular stakeholders should be fostered. This argument can best be illustrated by an example.

In 2016, a ProPublica report (Angwin et al., 2016) investigating a widely used tool for predicting recidivism risk found that the software was biased against Afro-American defendants, significantly overestimating their likelihood to re-offend (false positive error) while underestimating the risk of Caucasians to commit future crimes (false negative error). Northpointe (now Equivant), the company developing the COMPAS software, argued in reply that their test was racially neutral because the rate of accuracy – around 60 percent – was the same for both groups, leading the company to "strongly reject the conclusion that the COMPAS risk scales are racially biased against blacks" (Dieterich et al., 2016: 1). What is interesting here are not so much the technical details of the case,[10] but the fact that the dispute sparked a wider public conversation over algorithmic bias and motivated researchers to reflect and experiment on different notions of fairness and the inherent trade-offs between them (see, e.g., Kleinberg et al., 2017). Despite a growing body of technical literature, however, it is important to recognize that the development of fair algorithms will always involve choices between competing values that cannot be decided on purely technical grounds, but necessitate broader public deliberation.[11] As Wong (2019: 2) argues, the "'fairness' in algorithmic fairness should be conceptualized first and foremost as a *political* question and be resolved *politically*", meaning that the task will not merely be to optimize and improve relevant techniques for fair algorithms, but to "consider and accommodate diverse, conflicting interests in a society." (ibid.) Thus, especially in high-stake contexts such as criminal justice, decisions as to what counts as fair and what does not should

[10] Northpointe's answer to ProPublica's allegations comprises 37 pages to which the ProPublica journalists once again issued a detailed response.

[11] For reflections on the distributed structure of decision systems and related questions of responsibility, see chapter 5 by Katherine J. Strandburg in this volume.

not be made by a lone company or organization, but demand *some* form of democratic legitimation as well as the transparent communication of inevitable value trade-offs that have been made when designing the system.

Ultimately, then, the realization of more trustworthy AI environments would ask developers in critical contexts not only to consider and publicly communicate the limits of their design, but to acknowledge the limits of their competence to make certain design decisions in the first place. In sensitive areas, questions such as "What is an appropriate measure of fairness?" or "How to balance the trade-offs between explainability and accuracy?" may go well beyond what developers should be able to decide on their own, even if taking a measured approach and acting with best of intentions. Industry standards can help to alleviate these challenges to some extent, but the existence of a standard alone should not be seen as a carte blanche to go ahead without turning back. Northpointe, too, conformed to such standards, but this does not mean that their software does not pose a major ethical challenge that should have been democratically discussed not only *before* investigative data journalists broke the story and extensive media coverage forced a public response, but even *before* the software was initially deployed. To be truly trustworthy, it would not have been sufficient to merely inform users about the nature and characteristics of the software, even if such communication efforts would disclose information regarding the limitations and potential shortcomings of the risk assessment system. Given the high-risk use case, the internal decision to opt for a specific fairness measure and offer the product on the market should not have been made as the company should have recognized that it was not really their decision to make: the nature and scope of the decision fell well outside the companies competence.

The case of COMPAS combines a number of characteristics that would have required a much more thorough assessment before being allowed on the market. Not only may its usage in U.S. courts severely impact the future lives of those affected by the software, the defendants also had no way to opt out or contest the predictions. More generally, the following rule of thumb may be useful to decide on the level of care, oversight and auditing needed for trustworthy AI system: the broader and deeper the potential for detrimental impact, the more vulnerable those affected and the more unavoidable the system – either due to its usage in public administration or due to (quasi-)monopolies – the higher the requirements and controls for AI systems must be. With respect to the COMPAS case, trustworthy AI would thus not only imply that "[i]n situations in which no ethically acceptable trade-offs can be identified, the development, deployment and use of the AI system should not proceed in that form" (AI HLEG, 2019: 13), but also require considerable awareness amongst practitioners that in certain situations the decision of what constitutes an *acceptable* trade-off is simply not theirs to make, but must in fact be negotiated in a different – and in this case public – kind of deliberative setting. Which leads us to another issue.

3.3 The Moral Requirements for Trustworthy AI

As stated, trustworthiness does not only require that the trustee is *competent* to satisfy the trustor's demands, it also requires a specific *moral* component, namely being *trust-responsive* and *truthful*. With regard to trustworthy AI, this moral requirement faces obstacles on two separate fronts: on a *technical level*, because of the ways AI systems function, and on an *organizational level*, because of various operational and market incentives. We shall discuss both dimensions in turn.

First, when considering AI systems on a purely technical level, the argument can be made that trust-responsiveness may present too strong a requirement as it mandates the trustee to complete the tasks not (only) for her own values and interests but for the values and interests of her trustor.

Trustworthy AI, therefore, must find means to be responsive to the users' values and interests and demonstrate this responsiveness to them. However, most AI systems only treat users as members of certain classes or categories based upon the data points these users left behind. Consequently, individual values and interests are only seen as a form of probabilistic affiliation, e.g. the likelihood that a user values and enjoys what similar users value and enjoy. Thus, if trustworthiness requires trust-responsiveness on a personal level, the statistical basis upon which AI technologies operate may impose an insurmountable constraint for AI systems to be trustworthy. Given these fundamental limitations it may be sensible to weaken the requirement and redefine the necessary level of trust-responsiveness as being responsive to *group* values and interests in lieu of *personal* values and interests, and require AI systems to be trust-responsive only to those who are significantly impacted by them. This refinement is sensible as those who are significantly impacted by AI systems are in fact most vulnerable to them and thus truly require the AI systems to be trustworthy. Yet, even with this refinement, trustworthy AI still needs to find means to communicate to its users that it is acting on behalf of their values and interests in order to demonstrate itself to be trust-responsive.

The requirement of truthfulness implies that trustworthy AI must inform its users about its goals and workings. In this respect, the ethics guidelines include requirements for both transparency and accountability measures, involving demands such as the communication of an AI system's capabilities and limitations, the identification of relevant goals and interests, and the reporting of potential negative impacts, including accessible mechanisms of redress should adverse impacts occur. Some of these requirements may be easier to fulfill than others. For instance, the guidelines require that AI systems must be identifiable as such, since humans have the right to be informed that they are in fact interacting with an AI. Disclosures of this kind may be relatively easy to accomplish as a clearly visible disclaimer or short announcement in case of verbal interaction could suffice to inform users that they are now interacting with software and silicon rather than human brain cells. Disclosure of the specific goals and workings of an AI system, however, may be more challenging because they may be in conflict with a variety of organizational interests, which brings us to a second crucial point to be made.

When designing and deploying an AI system, it may be in the best interest of a company or organization not to disclose certain information for a variety of reasons. For instance, companies may wish to treat certain aspects of their AI system as confidential trade secrets to protect their investments and secure an economic advantage over competitors. They may also argue, as Google has with regard to its search engine, that the disclosure of too much information could make it easier for others to manipulate or 'game' the system, thereby reducing the quality and relevance of the service. Being forthcoming about the goals of an AI system may also conflict with business interests as the objectives are often much more mundane than marketing language would suggest. Facebook's official mission statement may hold that the platform strives "to give people power to build community and bring the world closer together" (Facebook, 2019), but at the end of the day it is a publicly traded company that is responsible to its shareholders in a competitive market. As Facebook CEO Mark Zuckerberg bluntly replied when asked by congress about the company's business model: "Senator, we run ads" (Washington Post, 2018).

The significant influence of financial interests on research has long been recognized in research ethics and bioethics (see, e.g., Resnik, 2007). Indeed, conflicts of interests may encourage the fabrication and falsification of research data as well as failure to disclose relevant information regarding research agendas and interests (ibid.). It is conceivable that financial interests have the same kind of influence in the AI context. Selective disclosure of business practices as in the case of the Cambridge Analytica (see Cadwalladr, 2018) or non-disclosure of controversial projects as in the case of Google's Dragonfly project (see Google Employees against Dragonfly, 2018) may serve as illustrative examples. To clarify, we do not wish to argue that the sheer existence of business considerations is immoral, but that such considerations may reduce efforts of information disclosure and thus create a challenge for AI companies to be truthful and, in extension, trustworthy.

4. Conclusion

Based upon our philosophical analyses on trust and trustworthiness, we can conclude that their moral and epistemic implications bring about four challenges for trustworthy AI. Using the human-centered terminology from philosophical accounts of trust and trustworthiness,[12] AI systems must, first, be *competent*, i.e. they must be designed and function properly and reliably. Second, AI systems must always also possess *adequate epistemic self-assessment*, i.e. they should monitor and display their premises, shortcomings and limitations. Third, AI systems must be *responsive* to the interests of the human trustors. While this may not be achievable in principle on an individual basis, we may require at least *responsiveness to the most vulnerable groups of users*. Fourth, the requirement of *truthfulness* demands honesty in regards to the goals and inner workings of an AI system in so far as it affects the users treatment through such systems.

We have shown that there are practical obstacles to achieving these requirements but would argue that there are ways to successfully deal with them. In order to do so, however, one must work toward facilitating a *trustworthy AI culture* or, as the *Ethics Guidelines for Trustworthy AI* poignantly stress, "ensuring Trustworthy AI requires us to build and maintain an ethical culture and mind-set through public debate, education and practical learning" (AI HLEG, 2019: 9). But while we would agree that the goal should be "to create a culture of 'Trustworthy AI for Europe'" (ibid.: 39) that respects and protects foundational values, we are much more skeptical about recent efforts to sell trustworthy AI as a ready-made label or brand (see, e.g., EC, 2019), simply because the moral requirements for trustworthy AI systems can hardly be satisfied through a standardized certification process, but must be *cultivated* as part of a much more comprehensive dialogical approach. What Europe should esteem to build is an *ethos* of trustworthy AI that generates safe and reliable AI products. The development of such products should indeed follow ethical guidelines such as the ones of the High-Level Expert Group outlined and discussed in this paper, but *trust* should ultimately only be extended to the democratic and political culture surrounding these products. Transparency or accessible mechanisms for redress, for instance, is not something that a company or organization should have the discretion to offer, but something that users – in due consideration of sectorial or contextual limitations – should have the ability to demand. As more and more countries issue guidelines that emphasize the need for trustworthy AI systems, it will only be the morally competent AI culture that can claim any credibility to truly trustworthy conduct.

[12] For a contrasting perspective, see chapter 6 by Federico Cabitzo in this volume, who argues for a shift from trying to make *AI human-centered* to pursuing an *AI-decentered humanity*.

5. References

AI HLEG (High-Level Expert Group on Artificial Intelligence) (2019). *Ethics Guidelines for Trustworthy AI*. [online] Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Ananny, M. (2016). Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Science, Technology, & Human Values*, 41(1), pp. 93–117.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine Bias. *ProPublica*. [online] Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

BAAI (Beijing Academy of Artificial Intelligence) (2019). *Beijing AI Principles*. [online] Available at: https://www.baai.ac.cn/blog/beijing-ai-principles.

Baier, A. (1986). Trust and Antitrust. *Ethics*, 96(2), pp. 231–260.

Burrell, J. (2016). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society*, 3(1), pp. 1-2.

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: *Proceedings of Machine Learning Research*, 81, pp. 1–15.

Cadwalladr, C. (2018). 'I Made Steve Bannon's Psychological Warfare Tool': Meet the Data War Whistleblower. *The Guardian*. [online] Available at: https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump.

Castro, D., McLaughlin, M., and Chivot, E. (2019). Who Is Winning the AI Race: China, the EU or the United States? *Center for Data Innovation*. [online] Available at: https://www.datainnovation.org/2019/08/who-is-winning-the-ai-race-china-the-eu-or-the-united-states/.

Dastin, J. (2018). Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. *Reuters*. [online] Available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Davies, J. (2019). Europe Publishes Stance on AI Ethics, but Don't Expect Much. *Telecoms.com.* [online] Available at: http://telecoms.com/498190/europe-publishes-stance-on-ai-ethics-but-dont-expect-much/.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*. [online] Available at: http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

Eubanks, V. (2018). *Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

European Commission (EC) (2019). *Building Trust in Human-Centric Artificial Intelligence*. COM(2019)168. [online] Available at: https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence.

European Commission (EC) (2018). *Artificial Intelligence for Europe*. COM(2018)237. [online] Available at: https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe.

Facebook (2019). Terms of Service. [online] Available at: https://www.facebook.com/legal/terms?ref=pf.

Google Employees Against Dragonfly (2018). We Are Google Employees. Google Must Drop Dragonfly. *Medium*. [online] Available at: https://medium.com/@googlersagainstdragonfly/we-are-google-employees-google-must-drop-dragonfly-4c8a30c5e5eb.

Gunkel, D. (2018). The Other Question: Can and Should Robots Have Rights? *Ethics and Information Technology*, 20(2), pp. 87–99.

Hardwig, J. (1991). The Role of Trust in Knowledge. *Journal of Philosophy*, 88(12), pp. 693-708.

Hardwig, J. (1994). Toward an Ethics of Expertise. In: Daniel Wueste (Ed.), *Professional Ethics and Social Responsibility*. Lanham, MD.: Rowman & Littlefield, pp. 83-101.

Hidvegi, F. and Leufer, D. (2019). Laying Down the Law on AI: Ethics Done, Now the EU Must Focus on Human Rights. *Access Now*. [online] Available at: https://www.accessnow.org/laying-down-the-law-on-ai-ethics-done-now-the-eu-must-focus-on-human-rights/.

Jones, K. (1996). Trust as an Affective Attitude. *Ethics*, 107(1), pp. 4–25.

Jones, K. (2012). Trustworthiness. *Ethics*, 123(1), pp. 61–85.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). *Inherent Trade-Offs in the Fair Determination of Risk Scores*. 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), 43, pp. 1-23.

Lamberth, M. (2019). US' AI Ethics Debate: Overcoming Barriers in Government and Tech Sector. *RSIS Commentary*, 129. [online] Available at: https://think-asia.org/bitstream/handle/11540/10492/CO19129.pdf?sequence=1.

Loritz, M. (2019). Europe's Call for Human-centric, Trustworthy AI Will Create More Opportunities for Startups. *EU-Startups*. [online] Available at: https://www.eu-startups.com/2019/07/europes-call-for-human-centric-trustworthy-ai-will-create-more-opportunities-for-startups/.

Marchant, G. (2019). "Soft Law" Governance of Artificial Intelligence. *AI Pulse*. [online] Available at: https://aipulse.org/soft-law-governance-of-artificial-intelligence/#easy-footnote-bottom-28-132.

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

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.

OECD (Organisation for Economic Co-operation and Development) (2019). *Recommendation of the Council on Artificial Intelligence*. [online] Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

O'Neil, C. (2016). *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishers.

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

OSTP (White House Office of Science and Technology Policy) (2019). *Accelerating America's Leadership in Artificial Intelligence*. [online] Available at: https://www.whitehouse.gov/articles/accelerating-americas-leadership-in-artificial-intelligence/.

Resnik, D.B. (2007). *The Price of Truth. How Money Affects the Norms of Science*. New York: Oxford University Press.

Ruokonen, F. (2013). Trust, Trustworthiness, and Responsibility. In: P. Mäkelä and C. Townley (Eds.), *Trust: Analytic and Applied* Perspectives. Leiden: Brill, pp. 1–14.

Samuel, S. (2019). How Do You Make Sure AI Is Trustworthy? The EU Wrote a Checklist. *Vox*. [online] Available at: https://www.vox.com/future-perfect/2019/4/9/18303539/ai-eu-trustworthy-guidelines.

Simon, J. (2013). Trust. In: D. Pritchard (Ed.), *Oxford Bibliographies in Philosophy*. New York: Oxford University Press. [online] Available at: https://www.oxfordbibliographies.com/view/document/obo-9780195396577/obo-9780195396577-0157.xml.

Spielkamp, M. (2019). *Automating Society. Taking Stock of Automated Decision-Making in the EU*. A report by AlgorithmWatch in cooperation with Bertelsmann Stiftung, supported by the Open Society Foundations. [online] Available at: https://algorithmwatch.org/wp-content/uploads/2019/01/Automating_Society_Report_2019.pdf.

Tallant, J. (2019). You Can Trust the Ladder, But You Shouldn't. *Theoria*, 85(2), pp. 102–118.

Wong, P.-H. (2019). Democratizing Algorithmic Fairness. *Philosophy and Technology*, 33, pp. 225-244. https://doi.org/10.1007/s13347-019-00355-w.

Washington Post (2018). *Transcript of Mark Zuckerberg's Senate Hearing*. [online] Available at: https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/.