AI and Ethics (2023) 3:735–744
https://doi.org/10.1007/s43681-022-00200-5

ORIGINAL RESEARCH

Trust and trustworthiness in AI ethics

Karoline Reinhardt[1,2]

Received: 27 April 2022 / Accepted: 23 July 2022 / Published online: 26 September 2022

Abstract
Due to the extensive progress of research in artificial intelligence (AI) as well as its deployment and application, the public debate on AI systems has also gained momentum in recent years. With the publication of the *Ethics Guidelines for Trustworthy AI* (2019), notions of trust and trustworthiness gained particular attention within AI ethics-debates; despite an apparent consensus that AI should be trustworthy, it is less clear what trust and trustworthiness entail in the field of AI. In this paper, I give a detailed overview on the notion of trust employed in AI Ethics Guidelines thus far. Based on that, I assess their overlaps and their omissions from the perspective of practical philosophy. I argue that, currently, AI ethics tends to overload the notion of trustworthiness. It thus runs the risk of becoming a buzzword that cannot be operationalized into a working concept for AI research. What is needed, however, is an approach that is also informed with findings of the research on trust in other fields, for instance, in social sciences and humanities, especially in the field of practical philosophy. This paper is intended as a step in this direction.

Keywords AI ethics · Fairness · Guidelines · Transparency · Trust · Trustworthiness

1 Introduction

Due to the extensive progress of research in Artificial Intelligence (AI) as well as its deployment and application, the public debate on AI systems has also gained significant momentum in recent years. It has become increasingly clear that computerized algorithms and AI pose not only technical but also profound ethical challenges [1–24]. We have thus witnessed the publication of a number of guidelines issued by research institutions, private companies and political bodies that deal with various aspects and fields of AI ethics such as privacy, accountability, non-discrimination and fairness. With the publication of the *Ethics Guidelines for Trustworthy AI* by the Independent High-Level Expert Group on Artificial Intelligence (2019), the notions of trust and trustworthiness gained particular attention within AI ethics-debates with regard to AI governance; the term trustworthy AI (TAI) has since been widely adopted by the AI research community, [25–39] public sector organizations, and political bodies issuing AI ethics guidelines. Despite an apparent consensus that AI should be trustworthy, it is less clear what trust and trustworthiness entail in the field of AI and what ethical standards, technical requirements and practices are needed for the realization of TAI.

In this paper, I give a detailed overview on the notion of trust employed in AI Ethics Guidelines. Such an overview of the current state of AI ethics is necessary to be able to define points for further research in the field of TAI. I focus on the following questions: to what extent and how is the term currently used in guidelines? What notions of "trust" and "trustworthiness" are prevalent in these guidelines? What political, social, and moral role is attributed to trust, there? Which concepts are associated with the term trustworthiness? (§2) Based on this overview I assess these findings regarding their overlaps and their omissions from the perspective of practical philosophy (§3). Practical Philosophy has a long-standing tradition in thinking about the notion of trust [32] in oneself [40–42] and in inter-personal, [43] social [40, 44–46] and political settings [47, 48] as well as in human/artefact interaction, that is for instance trust in technology [49–54]. I argue that currently AI ethics overloads the notion of trust and trustworthiness and turns

✉ Karoline Reinhardt
karoline.reinhardt@uni-tuebingen.de;
karoline.reinhardt@uni-passau.de

[1] International Center for Ethics in the Sciences and Humanities, Universität Tübingen, Tübingen, Germany

[2] Faculty of Arts and Humanities, Junior professorship for Applied Ethics, Universität Passau, Passau, Germany

it into an umbrella term for an inconclusive list of things deemed "good". Presently, "trustworthiness", thus, runs the risk of becoming a buzzword that cannot be operationalized into a working concept for AI research. On top of that, we can observe that the notion of "trust" deployed in AI research so far is mainly an instrumental notion. I, then, discuss whether the notions of trust and trustworthiness deployed in the guidelines are apt to capture the nature of the interaction between humans and AI systems that we want them to describe. What is needed is, as I argue, an approach to trust that is also informed with a normative foundation and elements emphasized in the research on trust in other fields, for instance, in social sciences and the humanities, especially in the field of practical philosophy. This paper is intended as a first step in this direction. In §4, I formulate points to consider for future research on TAI. In the final section I draw some more general conclusions on how we should conceptualize trust regarding AI and which mistakes we should avoid (§5).

2 Trust and trustworthiness in AI ethics: an evaluation of guidelines

2.1 Corpus of documents

The past years have seen a rise in publications of AI ethics guidelines and frameworks with new documents being disseminated every month. In the past two years, a few helpful overviews of ethics guidelines have also been published [9, 29, 55–57]. For my analysis, I combined the samples of Jobin et al. [55], Zeng et al. [56], Fjeld et al. [57], Hagendorff [9] and Thiebes et al. [29], and added documents that were either published after the publication of the five overview articles or excluded for other reasons not relevant for this survey [9, 57–63].[1] Different from the other overviews that cover with one exception [29] the entire landscape of AI ethics, I focus on the notions of trust and trustworthiness and key concepts associated with these two notions in the guidelines. Some view trust and trustworthiness as a central principle, [55] others merely mention it [57] and yet others do not discuss this topic at all [9, 56]. Thiebes et al. [29] give an overview on eight trustworthy AI guidelines[2] Unlike them, I do not start with a fixed definition of trustworthiness, but first assess how the term is used.

2.2 Trust, trustors and trustees?

Trust is, in the guidelines I analyzed, generally perceived as something "good". Only few guidelines warn against blind [64], excessive trust [59, 65], being overly trusting [66] and "abuse of the trust that is artificially built between humans and machines"[67]. Some suggest that there might be "appropriate" or "correct" levels of trust [66, 67]. Most guidelines refer to trust, however, as something to be advanced [68], built [66, 69–78], created, [79] cultivated, [80] earned [73, 76, 81], elevated [66], enabled [82], established [69, 80, 83], gained [83–85], fostered [63, 69, 74, 77, 86, 87], improved [71], increased [66, 67, 83, 87], inspired [74], maintained [72, 78, 85, 88], promoted [63, 69, 74, 75, 87], reinforced [75], and upheld [63]. Rarely it is mentioned, however, that trust can also decrease [89], erode [62, 89, 90], be undermined [87], or even lost [76] and in some cases has to be restored [76].

The guidelines differ in the envisioned addressee of trust building; sometimes it is the general public or society as a whole that is addressed and the aim is to build public or social trust [64, 67, 71–73, 75, 81, 85, 87, 90, 91]. Corporate ethics guidelines as well as guidelines by business associations, unsurprisingly, tend to emphasize the trust of clients, consumers, customers, and users, but other organizations do as well [66–69, 71, 74, 76, 77, 81–83, 85, 92]. Less mentioned is human trust or trust between humans and technology [66, 89, 93], machines [79] or robots [67, 84]. The trust of citizens [72, 75], individuals or people (irrespective of their societal role) [69, 71], employees [71], workforce trust [94] or stakeholders [71] are only occasionally mentioned. In guidelines on AI use in the health and care sector clinicians, physicians, and practitioners also play a role [72, 83].

The proposed answer to the question who or what is to be trusted also diverges among guidelines. In the *Asilomar AI Principles*, we read about trust and transparency among AI researchers and developers [95].[3] Others see developers and

[1] A few words on the limitations of the corpus of documents are warranted: Guidelines and frameworks are grey literature and therefore not indexed in scholarly databases. Their retrieval is thus less replicable than searches for academic literature: some of the online sources referred to in Jobin et al. [57] and Hagendorff [9] could no longer be retrieved and were, thus, removed from the corpus. Occasionally, older versions of documents were replaced by updated more recent ones. The issue of a possible bias in searching for documents that Jobin et al. [57] have already discussed was mitigated by including studies and overviews from various authors. In the corpus of documents we probably still face a bias towards languages using Latin script and in particular towards results that were written or translated into English as it was already present in the five overviews that I used as a basis. In the selection of documents, no distinction has been made between "framework", "guideline", "principles", "recommendation" and "white paper". There is of course an extensive overlap regarding the corpus of documents between the five overviews: More than 120 documents were considered. After removing duplicates roughly 100 documents remained and were further analyzed.

[2] They treat "beneficial", "ethical", "responsible" and "trustworthy" AI as roughly synonymous which leads to some conceptual blurring.

[3] *Ethically Aligned Design* stresses that an "atmosphere of trust" among AI researchers is needed for empowering AI researchers to share their worries and live up to the standards of professional ethics

designers also, but not solely, as the appropriate addressees of calls for trustworthiness [105].

2.3 What makes AI trustworthy?

In the guidelines, trustworthiness is linked to a whole range of different principles; transparency does play a major role and is linked to trust and trustworthiness in a number of guidelines [60, 63, 66–69, 73–76, 79, 81, 83, 86, 87, 92, 94, 96–104]. Trustworthiness is, however, also linked to reliability [60, 87, 93] and robustness [63, 87, 97] as well as safety and security [59, 60, 63, 67, 68, 72, 84, 91, 94, 97, 105].

Traceability [96] and verifiability [59, 66, 85] are sometimes mentioned in relation to trustworthiness [106, 107]. Other guidelines require TAI to be understandable [72, 76, 94, 101], interpretable [71] and explainable [63, 67, 71, 72, 74, 81, 85, 86, 94, 96]. IBM explicitly states: "we don't blindly trust those who can't explain their reasoning" [102]. And it is further mentioned that opacity might lead to a lack of trust [69]. The notion of predictability of an AI systems is also sometimes invoked when referring to trustworthiness [96].

In some guidelines the role of accountability for TAI is emphasized [62, 63, 74, 91, 94, 96, 97, 104]. Some also state that questions of liability need to be addressed for AI to be trustworthy as well as potential misuse scenarios [62, 91]. Monitoring and evaluation processes [62, 74, 86, 108] as well as auditing procedures should be in place for TAI [74, 96]. Human oversight [64, 97] as well as oversight by regulators [105] is sometimes one of the key requirements. The importance of compliance with norms and standards is highlighted in some guidelines [83, 98], or, in the terminology of the HLEG: TAI should respect all applicable laws and regulations and, thus, be lawful [97]. Some guidelines stress that TAI should be aligned with moral values and ethical principles [66, 69]. Others state that "an appropriate regulatory approach that reduces accidents can increase public trust" [87]. Rarely it is mentioned that accreditation, approval and certification schemes and the formulation of guidelines can help establish trust [69, 72, 83].

Also mentioned is the principle of avoiding harm [74, 86, 104, 109]. According to some guidelines, TAI should not only avoid harm, but also promote good or be beneficial to people [66, 74]. What this good or benefits consists in, however, is rarely defined [74]. [4] Sometimes it is stated that AI systems should serve the needs of the user [76], or even enhance "environmental and societal well-being" [97] and the "good of humanity, individuals, societies and the environment and ecosystems" [86]. Other guidelines emphasize "economic growth and women's economic empowerment" [91], or, "inclusive growth, sustainable development and well-being" [63]. Taking the public interest into account is also mentioned [71, 74, 83, 94, 104]. In guidelines focusing on the health care sector, success rate is also invoked as a variable for determining an algorithm's trustworthiness [83].

Diversity, non-discrimination and fairness are further principles linked to TAI [63, 69, 74, 83, 86, 87, 94, 96, 97, 104]. Risks of biases need to be addressed [60, 62, 91, 94, 104]. Control over one's data [60, 69], in general [71] or the ability to guide machines as they learn [94] are also mentioned as important. Data security [62, 68, 91] and privacy protection [63, 69, 72, 75, 97] are yet further principles invoked in the guidelines in relation to TAI. TAI shall also foster human agency, [97] autonomy, [60, 97] human-centered values [63, 66, 74] and respect for human rights [104]. The relation of trust and consent is, though an important topic in the philosophy of trust [48], however, rarely explored in AI ethics guidelines [83, 86, 97]. Even less frequently, non-manipulation is referred to [104]. Maybe somewhat surprisingly the question of who developed an AI [83], who trained it and what data was used during training [84], is rarely mentioned as relevant for trust.

Though the connection between trust and a variety of ethical principles is made in the guidelines, no single principle is linked to trust as making AI trustworthy throughout the entire corpus of documents. The conceptualization of trust in general and the definition of what makes AI trustworthy are thus so far inconclusive.

3 Assessing AI ethics guidelines

3.1 Conceptual overlaps between the guidelines

In what follows, I want to discuss important overlaps between the guidelines. I will focus on five main points: firstly, the guidelines that refer to trust at all view building trust dominantly as a "fundamental requirement for ethical governance" [55] and as something "good" that shall be promoted. Whereas lack of trust is dominantly seen as something to be overcome [71]. Ambivalences regarding trust are rarely discussed. Some hint at the fact that blind trust may be problematic [64] and "that people are unduly trusting of autonomous mobile robots", [89] but that trust is a rather ambivalent concept [32] is rarely mentioned [89].

Secondly and strikingly, most guidelines are based on an instrumental understanding of trust: trust is described as something that is a precondition to achieve other things, like the benefits connected to AI or to realize AI's full potential

Footnote 3 (continued)

not only in severe cases of whistleblowing but throughout the entire process of research and deployment [68].

[4] PDPC, for instance, equates avoiding harm and beneficence [76].

for society. To give a few examples: in *Artificial Intelligence: Australia's Ethics Framework* issued by the Australian Governments Department of Industry Innovation and Science [73], AI is described as having "the potential to increase our well-being; lift our economy; improve society by, for instance, making it more inclusive; and help the environment by using the planet's resources more dito" [75]. To achieve these benefits, however, "it will be important for citizens to have trust in the AI applications" (ibid.). Thus, though described predominantly as something worth achieving trust is not perceived as an intrinsic value.

Thirdly, in the guidelines, dimensions of interpersonal concepts of trust, institutional and social trust and trust in technology are lumped together. All of them certainly play a role and should play a role when we discuss TAI. In what way and to what extent they are and ought to be relevant, however, as well as the question of what aspects of them are desirable in liberal democratic societies regarding TAI, still needs to be worked out and laid out more precisely.

Fourthly, no single principle stands out as being mentioned in all the guidelines as making an AI system trustworthy. As noted above, in the guidelines, "trustworthiness" is linked to a whole range of different principles. One almost gets the impression that everything that is considered "good" is also supposed to inspire trust and on top of that regarded as a necessary precondition for AI systems to be trustworthy. We thus run the risk of turning trustworthiness into an umbrella term for all things in general considered "good". Or to put it more polemically, we seem to expect AI systems to fulfill all the principles that we think we ought to fulfill on an interpersonal, societal, and political level such as justice, non-discrimination, and reliable protection of human rights, and fail to do so. This, however, turns trustworthiness in a buzzword that is not applicable or operationalizable.

Related to this point is, fifthly, that possible trade-offs and conflicts between these various values and principles that are supposed to generate trust are rarely reflected and how they would play out with regard to an AI system's trustworthiness [60]. Transparency and privacy, for instance, are things that we cannot have both at the same time, at least not with regard to the same entity [2]. Other conflicts between principles will not occur on a conceptual level like the one between transparency and privacy, but rather with regard to their application and implementation. Both points combined, overloading the notion of trustworthiness, and not addressing possible contradictions, tend to turn TAI into an intellectual "land of plenty", a mythological or fictional place where everything is available at any time without conflicts.

3.2 Conceptual omissions regarding trust in the guidelines

In the following, I concentrate on five omissions that are of particular importance: Firstly, what is overlooked in most of the guidelines is that trust has to do with uncertainty [110] and with vulnerability [32, 33, 43, 111]. We only need to trust where there is uncertainty about the outcome of a given situation and that outcome puts us at risk.[5] What is more: the event we then trust to happen can be positive or negative: "One can also trust in the end of the world" [113]. The object of our trust or event that we trust to happen does not necessarily need to be beneficent. That trust is an ambivalent concept is an observation which is often overlooked in the guidelines.

Related to this point is, secondly, that trust is often a fallback position in situations that are too complex to understand or where the costs of establishing understanding outweigh the supposed gains of doing so: trust helps to reduce social complexity, as an influential sociological account of trust argues [44]. Under this perspective, increasing transparency that most guidelines view as conducive to trust-building actually decreases the need for trust by decreasing uncertainty [48].[6] One could say, according to this account of trust: where I have all the information and have understood all the inner workings of the AI system, I do not really need to trust it anymore, because then I simply know how it works.[7]

Thirdly, we can observe a certain one-sidedness in the guidelines regarding the idea of how trust is established. The focus is clearly on the side of those who have an interest in building trust. Trust is very strongly portrayed as something that one can bring about, that needs to be improved, maintained, earned and gained: the dominant envisioned actor of the trust game is the trustee. When reading the guidelines, it sometimes appears as if bringing about trust were entirely under control of the trustee. The role of the trustor is not sufficiently reflected. This goes hand in hand with overlooking the fact that the trust game is ultimately an open-end-game;

[5] In the extensive philosophical literature on trust the epistemic preconditions of trust play an important role s. Alcoff [112], Daukas [113], Fricker [114], Horsburgh [115].

[6] O'Neill [48] argues that some sorts of transparency might even lead to spreading mistrust.

[7] Of course, there are other aspects of developing and using an AI model that, even if I know everything about the system itself, are still based on trust. AI systems, as socio-technical systems, are involved in a number of different overlapping trust relationships. This is already present in many guidelines. See for instance, "Trust in the development, deployment and use of AI systems concerns not only the technology's inherent properties, but also the qualities of the socio-technical systems involving AI applications." [108] I thank an anonymous reviewer for pointing this out.

it does not suffice for establishing trust that something or someone is trustworthy. The person who is supposed to trust has to grant trust as well.

Fourthly, the dynamic and flexible aspects of trust building as well as trust withdrawal are not in focus in the guidelines.[8] The fact, as already indicated in the previous point, that trust is a two-way affair and that the condition of the person trusting can also play a major role in whether trust is established or not, is still too little discussed. Reading the guidelines, one sometimes gets the impression that trust could and would never be withdrawn once it has been gained. This, obviously, is not correct and needs to be addressed in TAI research.

The last apparent omission I want to mention here is of a less conceptual, but a more practical nature; although it is occasionally mentioned in the guidelines that TAI systems should benefit the environment and the economy, or at least not harm them, surprisingly, even in those guidelines that focus on customer and user trust, the conditions under which AI products are created hardly plays a role and they are rarely mentioned as a factor for increasing or decreasing people's trust in AI. An exception is for instance *Ethical, Social, and Political Challenges in Artificial Intelligence in Health* that raises the question "who developed it?" and "What kind of data was the AI trained on? If I am a member of a minority group, will the AI work well for me?" as central to the question of whether an AI is trustworthy [83]. What is also rarely mentioned as a factor for TAI is, what resources are needed to produce AI systems, what working conditions prevail, what resources they require when they are in operation and who finances their development [9, 93].[9] Also, whether an AI system could potentially be deployed for military purposes does not play a major role.[10]

[8] One exception is Ethically Aligned Design: This guideline explicitly states that trust is dynamic [68].

[9] Deutsche Telekom raises the question of supplier chains and of whom one should choose to not work with in "order to engender trust". [95]

[10] Exceptions are the Ethics Guideline for Trustworthy AI [108] and the Recommendations of the Council on AI of the OECD [64]. Ethically Aligned Design discusses autonomous weapon systems. However, the term trust is used here only in the context of the accountability problem with respect to trusted user authentication logs [68]. It is not a question of whether the possibility of military use has an influence on the trustworthiness of certain systems or not. IIIM's Ethics Policy [116] makes a point of discussing military use of AI, not in relation to its trustworthiness, though.

4 Closing the gaps: future research on the T in TAI

4.1 Ethics and TAI

What can ethics do about these shortcomings? Regina Ammicht Quinn describes four types und understandings of ethics: a merely *ornamental understanding of ethics* that views ethics as "the icing on the cake" when everything is done; an *instrumental understanding of ethics*, something that we need to be done with, a box to check on a form; a *substantial-instrumental understanding of ethics* that provides orientation, for instance in the form of guidelines [110]. The understanding of ethics she advocates, however, is a *non-instrumental understanding of ethics*; under this perspective, ethics asks to critically reflect on and evaluate our (often implicit) presuppositions and their moral acceptability.

From an ornamental perspective, we would be content with the new label TAI and leave it at that. From an instrumental perspective, we provide one box to check on a form before an AI system is disseminated: "Is it trustworthy? Check." From a substantial-instrumental perspective, we provide checklists for TAI that are more comprehensive: "Is it transparent? Is it robust? Is it reliable? Is it explainable?" From a non-instrumental perspective, however, we must do a lot more footwork. Trust is presumably the basis of successful individual and social coexistence, it is, however, not in itself "good" in any moral sense [110], but a highly ambivalent concept. An ethical perspective on TAI needs to take this into account: we have to talk about false trust, misused trust, the perils of trust and (productive) distrust. This is not to say that there are no points that connect the current guideline-understanding and a more philosophically informed understanding of trust. On the contrary, both converge in a crucial point: the observation that trust is of instrumental, not of intrinsic, value.

The question is which elements of a philosophically informed concept need to be present in such a notion of trust when it comes to AI governance for it to (a) portray trust relations correctly, and (b) be a viable and ethically sound concept for AI research. This is not the place to fully develop such a notion of trust. Nevertheless, I would like to outline a few aspects that can provide a basis for further explorations in the next section and point out points for further research derived from these brief considerations.

4.2 AI governance

Technologies are not developed in a societal vacuum, but are interwoven with the fabric of our social and political interactions. They are, as socio-technical systems, embedded

in societal and political contexts [97]. This is particularly true with regard to AI technologies and algorithmic decision making; algorithms based on machine learning already shape our lives and social interactions in profound ways [3]. Liberal democracies, now, take a certain stance when it comes to social arrangements: "Liberals are committed to a conception of freedom and of respect for the capacities and the agency of individual men and women, and these commitments generate a requirement that all aspects of the social should either be made acceptable or be capable of being made acceptable to every last individual" [117]. Following this perspective, technologies, their development and deployment as part of our social arrangements must also meet these requirements. Especially when they are as interwoven with the workings of our social and political institutions as many AI systems are today.

The commitment to respect for the capacities and the agency of individuals leads also to the idea of checks and balances that modern democracies are based on. Checks and balances are essentially institutionalized forms of distrust and serve to ensure that no branch of government or any other institution abuses its power [118]. Ultimately, they also serve to safeguard individual agency. For many theorists of democracy, distrust is at the root of the basic set up of modern democracies: "Liberalism, and then liberal democracy, emerged from the distrust of traditional political and clerical authorities [119]."

In addition, it is worth noting that guidelines are governance instruments; they are part of the control and regulation system of private and public institutions. This is particularly obvious when they are issued by political institutions, such as the above quoted guidelines by the EU Commission or the UNESO, but it also holds for guidelines issued by private companies. As governance instruments, they are not only written for developers but are part of a wider system of regulation and control that must be in line with the above-mentioned requirements on the acceptability of social arrangements, especially when technologies have profound impact on these arrangements as many AI systems do.

4.3 Points for further research

I will focus on three aspects: the overlooked ambivalence of trust, the problematic conflation of trust and trustworthiness, and the problem of conflict of principles in the guidelines:

Ambivalence of trust

As mentioned above, trust is a highly ambivalent concept; trust "is important, but it is also dangerous"[32]. It makes us vulnerable, as has been often pointed out in the literature on trust [120]. The ambivalence of trust is, however, often obscured in the guidelines thus far, and should be taken into account in further research on TAI. One might now wonder, why should we care about the ambivalent nature of trust in AI governance?[11] The ambivalence of trust has to be addressed in order not only to appropriately capture the nature of trust relations, but also because of its practical relevance; trust comes with a number of ethically relevant risks. The nature of trust is thus not only of interest for classroom discussions, but of high practical importance. When algorithmic decisions are as interwoven with the fabric of society as they are today and increasingly will be in the future, this generates the requirement that these aspects of our lives and the risks that come with them fulfill basic requirements of justification, as described above. Hence, we might, ultimately, not be in need for more trust in the application of AI systems, but for structures that institutionalize "distrust" [110], like binding standards, or mandatory auditing and monitoring. Here, moral, social, and political philosophy can help to further clarify matters.[12]

Problematic conflation of trust and trustworthiness

Trust and trustworthiness get easily conflated in the guidelines as well as in debates on TAI. From an ethical standpoint it is of utmost importance to keep them apart conceptually; ideally, we only trust things and people that are trustworthy. However, this is obviously not how trust works. People trust things and persons that are utterly unworthy of trust, and they do not trust things and persons that are utterly trustworthy; to perceive something or someone as trustworthy and for something or someone to be trustworthy are two different things [121–123]. This is also relevant for AI design. To give but one example: when AI systems behave like humans, for example when they give natural language recommendations or even have a voice, like some virtual assistant systems, people tend to trust them more easily. Voices and faces might make it easier to trust a certain entity, but they tell us nothing about the trustworthiness of that entity.

The disconnection of trusting and trustworthiness cuts both ways. Someone or something might be utterly trustworthy and still people won't trust it. However, in the guidelines, as in the wider debate on trustworthiness, this is often overlooked. Thiebes et al., for instance, "propose that AI is perceived as trustworthy by its users […] when it is developed, deployed, and used in ways that not only ensure its

[11] I thank an anonymous reviewer for pointing this out.

[12] One might want to object that it is unclear what good it would bring to address this issue in practical guidelines and it is surely correct that there would probably be no good in turning a guideline into an essay about the nature of trust. However, when set up properly, guidelines are the condensate of a longer deliberation process and the awareness of the ambivalent nature of trust should inform this deliberation process. I thank an anonymous reviewer for pointing this out to me.

compliance with all relevant laws and its robustness but especially its adherence to general ethical principles " [29] Maybe it would be perceived as trustworthy, maybe it would not. As discussed above, it does not suffice to be trustworthy to gain trust; trust must be granted. Overlooking this observation has profound consequences for how we conceptualize the interaction called "trusting". Additionally, it has practical implications:

Putting so much emphasis on designing TAI as a means for a wider adoption of AI systems might in the end of the day lead to a great disappointment on the side of developers. Much effort might be put in designing trustworthy AI and people might still not trust it, let alone adopt it. [13]Trust can be earned, but it also has to be granted. There are good reasons to design trustworthy AI as there are good reasons for many of the values and principles mentioned in the guidelines. But they might ultimately not lead to a wider adoption of AI systems, simply because trustworthiness does not automatically lead to trust in the way that many of the guidelines seem to assume. It might not fulfill the expectations regarding technology adoption. This is not to say that we should not care about the principles that AI systems endorse or violate, we just might need to provide a different reasoning, or different incentives from gaining trust. Structures that institutionalize "distrust" like binding standards can also provide strong incentives to act in certain ways. Related to this point, the call for more trust needs to be monitored closely; in some cases, it might stand in for an avoidance of strict hard law regulations. The dominant perspective in the corpus of documents is that building trust is a "fundamental requirement for ethical governance" [55]. In many cases, however, it might be the more ethical decision to call for a robust legal framework, not for more trust.

On a more conceptual note, accepting that people are free to make their own decisions when it comes to trust and that they should not be lured, tricked, or coerced into trusting no matter how advantageous that would be for others or for themselves, is part of the respect for the capacities and agency of people mentioned above. Taking up the virtual assistant example: making it easier to trust it might in some cases even make it less worthy of trust, because people should not be lured into trusting. Ultimately, this comes down to a relatively profound argument: taking the self-determination and autonomy of people serious involves accepting that trusting is a game with an open end. It is their choice to trust, or not to trust.

Conflict of principles

Finally, the guidelines thus far join conflicting principles regarding the foundation of trustworthiness. This is problematic because it leaves developers unclear as to which principle should be applied in case of conflict, which should be given priority in specific cases or how conflicting values should be weighed against each other. This makes room for arbitrariness and opens the door for cherry-picking, ultimately, putting the whole endeavor of well-founded trust in AI at risk because practitioners or users cannot be sure which part of the trustworthiness canon was applied to what extent to a system in question, and which aspects were left aside. In further research on TAI, thus, it has to be addressed how trade-offs and conflicts between principles are to be resolved. One possibility is to significantly downsize the list of principles mentioned and thus reducing or even eliminating conflict of principles. Another possibility would be to introduce lexical prioritization of the principles related to trustworthiness. Yet another approach might challenge the aptness of the notion of trustworthiness regarding AI altogether.[14]In any case, this decision should be well-grounded in not only pragmatic, but also ethically sound reasons.

5 Conclusions

Though notions of trust and trustworthiness have gained significant attention in AI research especially after the publication of Europe's High-Level Expert Group's *Ethics guidelines for trustworthy AI,* ethics guidelines referring to trust with regard to AI diverge substantively in no less than four main areas: (1) why trust is important or of value, (2) who or what the envisioned trustors and trustees are, (3) what trustworthiness is and entails, and (4) how it should be implemented technically. Further clarification on all four points is needed. Philosophy can help to conceptualize trust and trustworthiness. At the same time, it is of utmost importance not to turn TAI into an intellectual land of plenty: it should not be perceived as an umbrella term for everything that would be nice to have regarding AI systems, both from a technical as well as an ethical perspective. Furthermore, we need to discuss possible conflicts between the various principles associated with trustworthiness in more detail. Finally, we should also take the ambivalences and perils of trust into account. In further research, it might turn out that in the end what we need is not more trust in AI but rather institutionalized forms of distrust.

Acknowledgements The research for this paper was conducted as part of the research project "AITE – Artificial Intelligence, Trustworthiness and Explainability" funded by the Baden-Württemberg Foundation (program: Verantwortliche Künstliche Intelligenz). I want to thank the participants of the AITE-Colloquia and the IZEW-Colloquium for their

[13] The contingencies of technology adoption are yet another matter.

[14] I am not taking a stance here on which possibility is to be preferred.

helpful comments on earlier versions of this paper. Special thanks are due to research assistants Oduma Abelio, Sandra Dürr and Lukas Kurz for their support with retrieving the guidelines and their diligence in proofreading the manuscript.

Author contributions Not applicable.

Funding Open Access funding enabled and organized by Projekt DEAL. The research for this paper was conducted as part of the research project "AITE – Artificial Intelligence, Trustworthiness and Explainability" funded by the Baden-Württemberg Foundation.

Availability of data and materials Not applicable.

References

1. Bostrom, N., Yudkowsky, E.: The Ethics of Artificial Intelligence. In: K. Frankish, W. Ramsey (eds.) The Cambridge Handbook of Artificial Intelligence, pp. 316–334. Cambridge University Press (2014)
2. Mittelstadt, B.D., et al.: The ethics of algorithms: mapping the debate. Big Data Soc. **3**, 2053951716679679 (2016)
3. O'Neil, C.: Weapons of Math Destruction. Brodway Books, New York, US (2016)
4. Taddeo, M., Floridi, L.: How AI can be a force for good. Science **361**, 751–752 (2018)
5. Floridi, L.: Establishing the rules for building trustworthy AI. Nature Machine Intelligence **1**, 261–262 (2019)
6. Barocas, S., Selbst, A. D.: Big data's disparate impact. California Law Rev. **104**, 671–732 (2016). https://doi.org/10.15779/Z38BG31
7. Bozdag, E.: Bias in algorithmic filtering and personalization. Ethics Inform. Technol. **15**, 209–227 (2013)
8. Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM transactions on information systems. Media Cult. Commun. **14**(3), 330–347 (1996). https://doi.org/10.1145/230538.230561
9. Hagendorff, T.: Maschinelles Lernen und Diskriminierung: Probleme und Lösungsansätze. Österreichische Zeitschrift für Soziologie **44**, 53–66 (2019)
10. Heesen, J., Reinhardt, K., Schelenz, L.: Diskriminierung durch Algorithmen vermeiden. In: Bauer, G., Kechaja, M., Engelmann, S., Haug, L. (eds.) Diskriminierung und Antidiskriminierung. Transcript Verlag, Bielefeld (2021)
11. Veale, M., Binns, R.: Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. Big Data Soc. **4**(2), 1–17 (2017)
12. Zuiderveen Borgesius, Frederik. Discrimination, Artificial Intelligence, and Algorithmic Decision-Making. Strasbourg: Council of Europe (2018)
13. Costanza-Chock, S.: Design Justice: Community-led Practices to Build the Worlds We Need. The MIT Press, Cambridge (2020)
14. Friedman, B.: Introduction to the special issue: value sensitive design: charting the next decade. Ethics Inform. Technol. **23**, 1–3 (2021)
15. Verbeek, P.-P.: Materializing morality: design ethics and technological mediation. Sci. Technol. Human Values **31**, 361–380 (2006)
16. Hildebrandt, M.: Privacy as protection of the incomputable self. Theor. Inq. Law **20**, 83–121 (2019)
17. Nissenbaum, H.: Contextual integrity up and down the data food chain. Theor. Inq. Law **20**, 221–256 (2019)
18. Turilli, M., Floridi, L.: The ethics of information transparency. Ethics Inf. Technol. **11**, 105–112 (2009)
19. Hildebrandt, M.: Who needs stories if you can get the data? ISPs in the era of big number crunching. Philos Technol **24**, 371–390 (2011)
20. Leese, M.: The new profiling: algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. Secur. Dialogue **45**, 494–511 (2014)
21. Burrell, J.: How the machine 'Thinks.' Big Data Soc. **3**, 1–12 (2016). https://doi.org/10.1177/2053951715622512
22. Matthias, A.: The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. Ethics Inform. Technol. *6*, 175–183 (2004)
23. Miller, B., Record, I.: Justified belief in a digital age: on the epistemic implications of secret internet technologies. Episteme **10**, 117–134 (2013)
24. Reinhardt, K.: Diversity-sensitive social networks and responsibility. InfTars – Inform. Soc. **21**, 43–62 (2021). https://doi.org/10.22503/inftars.XXI.2021.2.4
25. Floridi, L.: Translating principles into practices of digital ethics: five risks of being unethical. Philos. Technol. **32**, 185–193 (2019)
26. Varshney, K.R.: Trustworthy machine learning and artificial intelligence. XRDS: crossroads. ACM Magaz. Stud. **25**, 26–29 (2019)
27. Harrison, T.M., Luna-Reyes, L.F.: Cultivating trustworthy artificial intelligence in digital government. Soc. Sci. Comput. Rev (2020). https://doi.org/10.1177/0894439320980122
28. Janssen, M., et al.: Data governance: organizing data for trustworthy artificial intelligence. Gov. Inf. Q. **37**, 101493 (2020)
29. Thiebes, S., Lins, S., Sunyaev, A.: Trustworthy artificial intelligence. Electron. Mark. (2020). https://doi.org/10.1007/s12525-020-00441-4
30. Shneiderman, B.: Human-centered artificial intelligence: reliable, safe & trustworthy. Int J Hum Computer Inter **36**, 495–504 (2020)
31. OECD: Trustworthy artificial intelligence (AI) in education: promises and challenges. https://www.oecd.org/education/trustworthy-artificial-intelligence-in-education.pdf (2020)
32. McLeod, C.: "Trust." The Stanford Encyclopedia of Philosophy. Edward N. Zalta (ed.). (2020) https://plato.stanford.edu/archives/fall2020/entries/trust/
33. Baier, A.: Trust and antitrust. Ethics **96**, 231–260 (1986)
34. Gambetta, D.: Trust. B. Blackwell, Oxford (1988)
35. Holton, R.: Deciding to trust, coming to believe. Australas. J. Philos. **72**, 63–76 (1994)
36. Jones, K.: Trust as an affective attitude. Ethics **107**, 4–25 (1996)
37. Lahno, B.: On the emotional character of trust. Ethic. Theory Moral Pract **4**, 171–189 (2001)
38. Uslaner, E.M.: The moral foundations of trust. Cambridge University Press, Cambridge (2002)
39. Hawley, K.: Trust, Distrust and Commitment. Noûs **48**, 1–20 (2014)

40. Govier, T.: Self-trust, autonomy, and self-esteem. Hypatia **8**, 99–120 (1993)
41. Lehrer, K.: Self-trust. Oxford University Press, Oxford (1997)
42. Foley, R.: Intellectual trust in oneself and others. Cambridge University Press, Cambridge (2001)
43. Potter, N. N.: "Interpersonal trust". The Routledge handbook of trust and philosophy. Abingdon, Routledge (2020)
44. Luhmann, N.: Vertrauen. Konstanz und München, Germany: UVK (1979)
45. Fukuyama, F.: Trust. Free Press, New York (1995)
46. Seligman, A.B.: The problem of trust. Princeton University Press, Princeton (1997)
47. Alfano, M., Nicole, H.: Trust in Institutions and Governance. In: J. Simon (ed.): The Routledge Handbook of Trust and Philosophy. Routledge, Abingdon, UK (2020)
48. O'Neill, O.: A question of trust: The BBC reith lectures 2002. Cambridge University Press, Cambridge (2002)
49. Ess, C.M.: Trust and new communication technologies: vicious circles, virtuous circles, possible futures. Knowl. Technol. Policy **23**, 287–305 (2010)
50. Taddeo, M.: Trust in technology: a distinctive and a problematic relation. Knowl. Technol. Policy **23**, 283–286 (2010)
51. Ess, Charles, M.: Trust and Information and Communication Technologies. In: J. Simon (ed.): The Routledge Handbook of Trust and Philosophy Abingdon, Routledge, pp. 405–420 (2020)
52. Coeckelbergh, M.: Can we trust robots? Ethics Inform. Technol. **14**, 53–60 (2012)
53. Grodzinsky, F., Keith, M., Marty, J Wolf.: Trust in artificial agents. In: J. Simon (ed.): The Routledge Handbook of Trust and Philosophy. Routledge, Abingdon, UK, pp. 298–312 (2020)
54. Sullins, J.P.: Trust in Robots. In: J. Simon (ed.): Routledge Handbook on Trust and Philosophy. Routledge, Abingdon, pp. 313–325 (2020)
55. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat Mach Intellig **1**, 389–399 (2019)
56. Zeng, Yi., Enmeng Lu., Cunqing, Huangfu.: Linking Artificial Intelligence Principles. arXiv preprint arXiv:1812.04814 (2018)
57. Fjeld, J., et al.: Principled Artificial intelligence. Berkman Klein Center Research Publication, Cambridge (2020)
58. EC White paper on artificial intelligence - A european approach to excellence and trust, COM(2020) 65 final. (2020). https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
59. Fraunhofer IAIS.: Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. (2019). https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_KI-Zertifizierung.pdf
60. Fraunhofer IAIS: Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. (2021). https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf
61. Plattform Lernende Systeme. Ethik-Briefing, München (2020)
62. Plattform Lernende Systeme.: Kritikalität von KI-Systemen in ihren jeweiligen Anwendungskontexten, München (2021)
63. OECD: Recommendation of the Council on Artificial Intelligence. (2021). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
64. Allistene.: Éthique de la recherche en robotique: Rapport n° 1 de la CERNA Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene. (2014). https://hal.inria.fr/hal-01086579/.
65. CNIL: How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence. (2017). https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf
66. IEEE: Ethically Aligned Design. IEEE Standards v1. (2016). https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf.
67. Rathenau Instituut.: Human Rights in the Robot Age. (2017). https://www.rathenau.nl/sites/default/files/2018-02/Human%20Rights%20in%20the%20Robot%20Age-Rathenau%20Instituut-2017.pdf.
68. National Science and Technology Council.: The National Artificial Intelligence. Research and Development Strategic Plan. (2016). https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf.
69. Information Commissioner's Office.: Big Data, Artificial Intelligence, Machine Learning and Data Protection. (2017). https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.
70. NITI: National Strategy for Artificial Intelligence. (2018). http://niti.gov.in/sites/default/files/2019-01/NationalStrategy-for-AI-Discussion-Paper.pdf.
71. Institute of Business Ethics: Business Ethics and Artificial Intelligence. (2018). https://www.ibe.org.uk/uploads/assets/5f167681-e05f-4fae-ae1bef7699625a0d/ibebriefing58businessethicsandartificialintelligence.pdf.
72. Department of Health and Social Care: Initial code of conduct for data-driven health and care technology. (2018). http://allcatsrgrey.org.uk/wp/download/informatics/www_gov_uk_government_publications_code_of_conduct_for_data_.pdf.
73. Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J., Hajkowicz, S.: Artificial Intelligence: Australia's Ethics Framework. Data 61 CSIRO. (2019). https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf.
74. PDPC – Personal Data Protection Commission Singapore. Discussion Paper on Artificial Intelligence (AI) and Personal Data (2018). https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/Discussion-Paper-on-AI-and-PD---050618.pdf.
75. Intel: Intel's AI Privacy Policy White Paper. Protecting Individuals 'Privacy and Data in the Artificial Intelligence World. Intel. (2018). https://blogs.intel.com/policy/files/2018/10/Intels-AI-Privacy-Policy-White-Paper-2018.pdf.
76. Microsoft: Responsible Bots: 10 Guidelines for Developers of Conversational AI. (2018). https://www.microsoft.com/en-us/research/publication/responsible-bots/.
77. Floridi, L., et al.: Ai4people - an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**, 689–707 (2018)
78. UK House of Lords: AI in the UK: ready, willing and able? (2018). https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf
79. IBM: Everyday Ethics for Artificial Intelligence. (2019). https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.
80. Report Montréal Declaration.: For a Responsible Development of Artificial Intelligence. (2018). https://monoskop.org/images/b/b2/Report_Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf.
81. IBM: IBM's Principles for Trust and Transparency. (2018). https://www.ibm.com/blogs/policy/wp-content/uploads/2018/06/IBM_Principles_SHORT.V4.3.pdf.
82. ITI - Information Technology Industry Council.: Information Technology Industry AI Policy Principles. (2017) https://www.itic.org/public-policy/ITIAIPolicyPrinciplesFINAL.pdf.
83. Future Advocacy: Ethical, Social, And Political Challenges of Artificial Intelligence in Health. (2018). https://wellcome.org/sites/default/files/ai-in-health-ethical-social-political-challenges.pdf.

84. EPSRC: Principles of Robotics. (2011). https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/.
85. Conference toward AI Network Society.: Draft AI R&D Guidelines for International Discussions. (2017). https://www.soumu.go.jp/main_content/000507517.pdf.
86. UNESCO. Recommendation on the Ethics of Artificial Intelligence. (2021). https://unesdoc.unesco.org/ark:/48223/pf0000379920#page=14
87. Executive Office of the President, USA: Guidance for Regulation of Artificial Intelligence Applications (2020). https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf?utm_source=morning_brew.
88. Sony: AI Engagement within Sony Group. (2018). https://www.sony.net/SonyInfo/csr_report/humanrights/AI_Engagement_within_Sony_Group.pdf.
89. Future of Humanity Institute: The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv preprint arXiv:1802.07228. (2018)
90. Ministry of Economic Affairs and Employment, Finland: Work in the Age of Artificial Intelligence. (2018). https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160980/TEMjul_21_2018_Work_in_the_age.pdf.
91. G7: Charlevoix: Common Vision for the Future of Artificial Intelligence. (2018). https://www.mofa.go.jp/files/000373837.pdf.
92. Deutsche Telekom: Guidelines for Artificial Intelligence (2018). https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366.
93. MI Garage: Ethics Framework. (2020). https://www.migarage.ai/ethics-framework/.
94. Accenture: Responsible AI: A Framework for Building Trust in Your AI Solutions. (2018). https://www.accenture.com/_acnmedia/PDF-92/Accenture-AFSResponsible-AI.pdf.
95. Future of Life Institute: Asilomar AI Principles. (2017). https://futureoflife.org/ai-principles/.
96. Beijing AI Principles: AI Principles. (2019). https://www.baai.ac.cn/news/beijing-ai-principles-en.html.
97. HLEG - High-Level Expert Group on AI: Ethics Guidelines for Trustworthy AI. (2019). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
98. Special Interest Group on Artificial Intelligence: Dutch Artificial Intelligence Manifesto. (2018). http://ii.tudelft.nl/bnvki/wp-content/uploads/2018/09/Dutch-AI-Manifesto.pdf.
99. IBM: Transparency and Trust in the Cognitive Era (2017). https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/.
100. IBM: Everyday Ethics for Artificial Intelligence (2018). https://www.ibm.com/design/ai/ethics/everyday-ethics/.
101. Internet Society: Artificial Intelligence and Machine Learning: Policy Paper. (2017). https://www.internetsociety.org/wp-content/uploads/2017/08/ISOC-AI-Policy-Paper_2017-04-27_0.pdf.
102. UNI Global: Top 10 Principles for Ethical AI. (2017). http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf.
103. RCP: Artificial Intelligence (AI) in Health. (2018). https://www.rcplondon.ac.uk/projects/outputs/artificial-intelligence-ai-health.
104. Unity Technologies: Introducing Unity's Guiding Principles for Ethical AI. (2018). https://blog.unity.com/technology/introducing-unitys-guiding-principles-for-ethical-ai.
105. Intel: Artificial Intelligence. The Public Policy Opportunity. (2017). https://blogs.intel.com/policy/files/2017/10/Intel-Artificial-Intelligence-Public-Policy-White-Paper-2017.pdf.
106. Sage: The Ethics of Code: Developing AI for Business with Five Core Principles (2017). https://www.sage.com/~/media/group/files/business-builders/business-builders-ethics-of-code.pdf.
107. The Royal Society: Machine learning: The Power and Promise of Computers that Learn by Example (2017). https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf.
108. AMA: Making Policy on Augmented Intelligence on Health Care. (2019). https://journalofethics.ama-assn.org/sites/journalofethics.ama-assn.org/files/2019-01/msoc1-1902_2.pdf.
109. German Federal Ministry of Transport and Digital Infrastructure. Report of the Ethics Commission Automated and Connected Driving. (2017). https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile
110. Ammicht Quinn, R.: Trust generating security generating trust. Behemoth A J. Civil. **8**, 109–125 (2015)
111. Warren, M.E.: Democracy and Trust. Cambridge University Press (1999)
112. Horsburgh, H.J.N.: The ethics of trust. Philos. Quart. **10**, 343–354 (1960)
113. Reemtsma, J. P.: Vertrauen und Gewalt: Versuch über eine besondere Konstellation der Moderne. Hamburg, Germany: Hamburger Edition (2013)
114. Simon, J.: The routledge handbook of trust and philosophy. Routledge, London (2020)
115. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors **46**(1), 50–80 (2004). https://doi.org/10.1518/hfes.46.1.50_30392
116. Icelandic Institute for Intelligent Machines: Ethics Policy. https://www.iiim.is/ethics-policy/3/.
117. Waldron, J.: Theoretical foundations of liberalism. Philos Quar **37**(147), 127–150 (1987)
118. Mühlfried, F.: Misstrauen: Vom Wert eines Unwerts. Stuttgart, Germany: Reclam, here p.30 (2019)
119. Warren, M.E.: Democracy & Trust. Cambridge University Press, Cambridge (1999), p. 1
120. Baier, A.: Trust. The tanner lectures on human values. Princeton University, Princeton (1991)
121. Alcoff, L.M.: On judging epistemic credibility: Is social identity relevant? Philos. Exch. **29**, 73–89 (1999)
122. Daukas, N.: Epistemic trust and social location. In Episteme **3**, 109–124 (2006)
123. Fricker, M.: Epistemic injustice. Oxford University Press, Oxford (2007)