AI & SOCIETY (2023) 38:1753–1761
https://doi.org/10.1007/s00146-022-01401-6

OPEN FORUM

The problem with trust: on the discursive commodification of trust in AI

Steffen Krüger[1] · **Christopher Wilson**[1]

Received: 5 April 2021 / Accepted: 25 January 2022 / Published online: 25 February 2022

Abstract
This commentary draws critical attention to the ongoing commodification of trust in policy and scholarly discourses of artificial intelligence (AI) and society. Based on an assessment of publications discussing the implementation of AI in governmental and private services, our findings indicate that this discursive trend towards commodification is driven by the need for a trusting population of service users to harvest data at scale and leads to the discursive construction of trust as an essential good on a par with data as raw material. This discursive commodification is marked by a decreasing emphasis on trust understood as the expected reliability of a trusted agent, and increased emphasis on instrumental and extractive framings of trust as a resource. This tendency, we argue, does an ultimate disservice to developers, users, and systems alike, insofar as it obscures the subtle mechanisms through which trust in AI systems might be built, making it less likely that it will be.

Keywords Artificial intelligence · Trust · Commodification · AI ethics · Discourse analysis

1 Introduction

Trust is all the rage in thinking about artificial intelligence (AI) and society. It dominates international principles and guidelines developed by the private sector (IBM 2018) and by multilateral organizations (OECD 2019). It takes center stage in national strategies for AI (Misuraca and Van Noordt 2020) and in the mandates for proposed national regulatory institutions (Mulgan 2016). The idea of trust grounds scholarly proposals for AI governance regimes (Janssen et al. 2020) and experimental tinkering in the programing of algorithms (Ribeiro et al. 2016; Toreini et al. 2020). It frames civil society calls for ethical AI (Balaram et al. 2018), is extolled as "essential" by social commentators (Larsson et al. 2019), and spurs the launch of boutique consultancies.[1] Indeed, a recent mapping of efforts to strengthen the governance of AI finds a proliferation of principles "from all sectors, including governments, corporations, and civil society," and notes that the "stated purpose of many of the principles is to forge trust: between governments and citizens, between corporations and consumers or users, and between AI researchers and the general public" (Cussins Newman 2020, p. 11).

Critical voices in this discourse are few and far between. After all, who would want to argue against the importance of trust and trustworthiness? Yet, some stakeholders have pointed towards the undue strain that is being put on the concept. A disillusioned member of the High-Level Expert Group responsible for developing EU ethics guidelines, for example, criticizes the notion of trustworthy AI as "conceptual nonsense":

> Machines are not trustworthy; only humans can be trustworthy (or untrustworthy). If, in the future, an untrustworthy corporation or government behaves unethically and possesses good, robust AI technology, this will enable more effective unethical behaviour (Metzinger 2019).

Warnings such as Metzinger's seem to go mostly unheard in a discourse that is dominated by enthusiasm for the societal and market benefits assumed to follow from AI technologies. Instead, the literature considered in this commentary suggests that the concept of trust has a near-universal utility. Trust in AI is imagined to help governments gain legitimacy

✉ Steffen Krüger
steffen.kruger@media.uio.no

Christopher Wilson
c.b.wilson@media.uio.no

[1] Department of Media and Communication, University of Oslo, Oslo, Norway

[1] See, for example, https://www.aisustainability.org/about-us/.

(Freuler and Iglesias 2018; Kuziemski and Misuraca 2020); to improve customer satisfaction with services (Al-Mushayt 2019; Berscheid and Roewer-Despres 2019; Matheny et al. 2019; Mehr 2017); help societies realize their potential for social good (Mikhaylov et al. 2018; Tomašev et al. 2020); and unlock market value (Chakravorti and Shankar 2017; Gordon Myers and Nejkov 2020; Orr and Davis 2020). Simultaneously, trust is widely applied as a quality seal for "good AI", and often collocated with "excellence", or "cutting-edge" technology, as in The European Commission's "White Paper on artificial intelligence: a European approach to *excellence and trust*" (European Commission, 2020, our emphasis).

A complementary strain of literature, meanwhile, warns of the risk that AI can fail to achieve its objectives when it is *not* trusted. Villani (2018), for example, states that "widespread distrust of AI on the part of the general public, […] in the long run is liable to curb its development and all the benefits it could bring" (116). Just as social media platforms wither away when they are not being fed with the interactions of a critical mass of people, AI will not develop sufficient intelligence—will remain dumb and unreliable—when left unused. Such "bad AI" can then lead into a vicious circle in which already lacking public trust is damaged further by the insufficient number of interactions it receives (Reisman et al. 2018).

These hopes and fears must be understood in light of the market forces which drive them. The emergence of AI technologies corresponds with the emergence of new markets, as indicated by stiff competition between private companies, and efforts by governments to strengthen national competitiveness (European Strategy for Data 2020; Misuraca and Van Noordt 2020). The coupling of strong incentives to advance AI systems with the inherent uncertainty surrounding AI's societal impact then leads to a polarized dynamic in which a skeptical public (Anderson et al. 2018) is matched by "AI cheerleaders" working to normalize socio-technical systems on behalf of private interests (Bourne 2019). The risk that this polarization can slide into paternalistic models of technological governance is very high indeed (Cardullo and Kitchin 2019; Oravec 2019) and should be taken seriously. As Cath (2018) notes when considering AI governance regimes

> …industry efforts [to develop best practices] are laudable, but it is important to position them in light of three important questions. First, who sets the agenda for AI governance? Second, what cultural logic is instantiated by that agenda and, third, who benefits from it? Answering these questions is important because it highlights the risks of letting industry drive the agenda and reveals blind spots in current research efforts (Cath 2018, pp. 3–4).

It is in this vein that we question the prominence of trust in the contemporary discourse on AI and society. To be clear, we do not assert that trust is in any way unimportant. Efforts that address widely held public concerns about AI's societal impacts are laudable, including efforts to build trust. We are concerned, however, that this near-universal attention to the concept obscures exactly the challenges that make trust in AI problematic in the first place.

This analysis is not an exploration of trust in AI or its conditions, whether they are present or ought to be, or whether trust is commodified in an economic sense of specific user interactions or market dynamics. Indeed, we are not concerned with trust per se, but the way in which the notion of trust is asserted and leveraged within a very specific discourse. This commentary is concerned with the commodification of trust as an important aspect, or moment, in the public discourse on AI and society, and the implications that moment might have for how AI is understood and implemented. Specifically, we are concerned by the increasing practical and instrumental attention to the general public's trust in AI as a precondition for generating the vast quantities of data required by machine-learning systems. Hence, trust becomes asserted as something that should be provided *blindly*, as a *resource to be extracted*, and as an *instrument for unlocking value*. Under these circumstances, the discursive commodification of trust is closely associated with the eagerness of private and government stakeholders to promote the implementation of AI systems in the name of trust, *while turning a blind eye to the question of whether those systems are actually deserving of people's trust*.

Our argument proceeds with a brief description of the theoretical background and methods that inform this commentary. This is followed by a third section outlining the relevant conceptions and levels of functioning of trust in the context of AI, before addressing the problems that can be associated with the existing perspectives. The fifth section concludes with an assessment of what trust has come to mean in the contemporary discourse and offers critical suggestions about how to counter the biases that this discourse produces to safeguard the common good and public well-being.

2 Theoretical background and method

Methodologically, what follows is best understood as a critical discourse analysis aligned with the understanding of socio-linguistic conventions as the expression of power relationships and power struggles (Fairclough 2010). In our application of the method, we depart from Michel Foucault's (1972) definition of discourse as relations between statements, groups of statements, and other events, even if these statements and events have no prior connection to one

another. In contradistinction to Foucault, however, we make room for the social effectiveness of non-discursive spheres in line with Norman Fairclough (1995) as well as the importance of a notion of subjectivity irreducible to social and discursive positions (e.g., Henriques et al. 1984; Carolan and Bell 2003). We believe that this is a productive approach to exploring the ways in which ideologies and hegemonic values are asserted and accepted in the discourses surrounding new technologies.

This approach is generally aligned with the tradition of critical social theory that understands these dynamics in terms of power and power relationships between groups and individuals (see Stoddart 2007). While the power relationships at play in this discourse are not the main focus of our analysis, this framing is particularly important in the context of AI's simultaneous opacity and ubiquity. We hope that our analysis can guide attention to the powerful interests that are driving the current discourse on AI governance, or to ask in Cath's (2018) language, who is setting the "agenda." This approach is also well suited to an analysis of the role of trust in such a discourse. Not only is trust widely recognized to be a discursive phenomenon, constructed and situated within social systems (Carolan and Bell 2003), but a critical discourse theory provides a productive framework for understanding how notions of trust weigh in on and influence governance paradigms, precisely because "trust-related and trust-bearing issues are central to our understanding of how the conduct of professional practices impacts on human relationships in social life" (Candlin and Chrichton 2013, p. 1). In this respect, our approach departs from the analysis often applied to the micro-level of communications in Information Systems research (Cukier et al. 2009), but remains focused at the macro-level of public discourse and how the social reality inscribed in it is indicative of the balances and imbalances of power in society. This also follows the theoretical orientation of critical policy studies to understand that the language used in public documents with a scholarly or policy orientation will play a key role in staking out what is politically possible and desirable in regard to AI governance as a policy problem (Mulderrig et al. 2019).

Furthermore, our theoretical orientation is reflected in our use of the term "commodification". There is a significant body of economic research conceptualizing trust as a commodity, which can be traced from Zucker's (1986) seminal work on the production of trust in macro-economic structures, to contemporary game-theoretical work on the commodification of trust in specific economic interactions (Dasgupta 1989). Recently, there has also been increased interest in the commodification of trust in regard to new technologies, and Bodø (2021) for example explores how blockchain technologies "transform trust and trustworthiness, a form of social capital, into a commodity, an industrially produced asset that can be quantified, traded, enclosed, and sanctioned" (p. 2678–2679). Our critical discourse analysis departs from this economic approach in favor of a Marxist understanding of commodification and a critique of neoliberal practices that leverage new technologies to assert and reinforce power relationships (Cardullo and Kitchin 2019). Unlike comparable critical readings, however (e.g., Arvanitakis 2007), we are attuned to this process at the level of discourse, rather than at that of individual's relationships or social realities.

Our sample for assessing the academic and policy discourse on AI and society is selective, and does not aspire to be strictly representative, inviting a critique that is often levied at critical discourse analysis (Breeze 2011). We nevertheless believe that a more partial review of literature is justified given the highly diffuse and fragmented nature of the discourse (Cussins Newman 2020) and the imprecise nature of trust as a concept. Our sample is nonetheless broad, composed of over 100 publications addressing AI, society, and trust. The sample includes academic journal articles ($n=65$), government strategies and whitepapers ($n=29$), analyses from Non-Governmental Organizations (NGOs) and think tanks ($n=10$), reports from industry consultancies ($n=2$), and journalistic opinion pieces and blog posts ($n=5$).

3 What we talk about when we talk about trust

For all its commonsensical appeal, trust is a surprisingly slippery concept, and as Hardin (2002) notes, there is "no Platonically essential notion of trust. Ordinary-language usages of the term trust are manifold and ill articulated" (p. xx). This ambiguity makes sense to a degree, if one understands trust as a social phenomenon, discursively constructed, bound by context, and continuously negotiated as a basis for social interaction (Candlin and Crichton 2013, p. 9). Ambiguity in this context is also what makes trust such a powerful social phenomenon, as a fundamental condition for complex social action and cooperation ("Trust Mak. Break. Coop. Relations" 1989), and underpinning the social construction of truth itself (Carolan and Bell 2003).

A common language understanding of trust will often emphasize the notion of trust cultivated through a process of learning from experience in interpersonal relations over time, and has been the focus of a significant strain of psychological research (e.g.: Berzoff 2011; Erikson 1963). This type of trust, situated in individual interactions and expectations, has been integrated with sociological conceptions of "system trust" in which trust is marked by increasing complexity and challenges to anticipating the behaviour of others, including complex social systems (Giddens 1990; Granovetter 1985; Luhmann 2017). Though sociological research has increasingly focused on the de-personalized

character of trust in such complex systems (Shapiro 1987), there is widespread agreement among scholars that individuals' trust—be it in other people, institutions, or systems—is always a matter of both prediction and normativity and is "always associated with expectations about the behaviour of others" (Candlin and Chrichton 2013, p. 2). This resonates with Hamm et al.'s (2016) cross-disciplinary review of what they call *institutional trust*, which emphasizes perceptions that institutions are protecting people's welfare as the primary antecedent of this type of trust.

For this commentary, it is not necessary to trace the messy debate about micro- and macro-level dynamics that contribute to building and eroding trust (e.g., Bachmann and Inkpen 2011). Indeed, there is every reason to believe that different types of trust are operating simultaneously, as has been demonstrated in experimental research on individuals' trust in online financial interactions (Pennington et al. 2003). To explore how the notion of trust is manifest in the discourse on AI and society, it is nevertheless useful to note that there are several ways in which trust might be cultivated in novel technologies like AI. While, in keeping with Hamm et al. (2016) and critics like Bodó (2021), trust might be earned with the exhibition of trustworthy behaviour, other studies indicate that, under certain conditions, people are willing to trust in emerging technologies even if these technologies have by no means earned their trust (e.g., Mazey and Wingreen 2017).

Furthermore, dynamics of trust in AI occur in imperfect information systems, and complex systems introduce multiple vectors for mistrust (Shapiro 1987). This has led to several proposals for how trust can be created and vested through proxy, including notions of trust mediators (Bodó 2021), data stewards (Janssen et al. 2020), or algorithmic social contracts to embed societal values in AI design (Rahwan 2017). Other research has emphasized the importance of institutional mechanisms to foster trust and mutual accountability between the different sectors involved in AI development (Brundage et al. 2020), and suggests that it might also be necessary to introduce macro-level systemic changes to cultivate trust at scale. This aligns with Steedman et al.'s (2020) exploration of trust in data-driven systems, "particularized solutions to generalized problems are unlikely to be effective. We need collective, ecosystem solutions, for example, better regulation of data-driven systems, in order for them to be perceived as more trustworthy" (p. 829).

Because trust is a matter of perception and predication, however, trust might also be cultivated without the performance of trustworthy behaviour. It is in this vein that Bourne and Edwards (2012) describe the discursive strategies deployed to cultivate system trust in the hedge fund industry. In the same vein, critics of AI systems lament the emphasis on public relations campaigns and marketing that attend to AI technologies and technology vendors (Bourne 2019; Cardullo and Kitchin 2019; Oravec 2019). For the present commentary and analysis, then the question becomes which of these responses is most prominent in the contemporary scholarly and policy discourse on AI and society.

4 Problems with trust

What becomes apparent from the discussion so far is that it confronts researchers and stakeholders with a challenge: To a significant degree, the object one is asked to invest one's trust in is still in the process of emergence. The very notion of AI is under public construction, both in the expert discourse reviewed here, and in the experiences and perspectives of the individuals that become AI users. Shifts and changes in sociocultural meanings affect the conditions for trust, and the global discourse sets the stage for how (and if) those conditions will be manifest. It is from this constructionist vantage point, which conceives of objects as being shaped by the ways in which they are understood, that we want to draw critical attention to the meanings that we see as emerging from the recent literature, where trust is frequently approximated to *blind trust*, to a *resource*, or an *instrument*.

4.1 Trust is blind

"[W]hether or not we are comfortable with AI may already be moot", states the World Economic Forum's report assessing the risks of AI systems (World Economic Forum 2017, p. 48). Its dry gesturing toward the inevitability of the implementation of AI illustrates just how much the question of trust in these systems is interwoven with the power relationships that support and are reinforced by them (Doteveryone 2018; Steedman et al. 2020). These power relations come yet clearer to the fore when the report explains that "To ensure that AI stays within the boundaries that we set for it, we must continue to grapple with building trust in systems that will transform our social, political and business environments, make decisions for us, and become an indispensable faculty for interpreting the world around us" (51). To taper the gist of this argument a little: *Since AI already is a social reality, we better find ways to trust it.*

Indeed, this rhetoric takes on a coercive tenor in some articulations, as for example in Polonski's (2018) assertion that trust in AI systems is necessary in order for users to avoid being left behind:

> [G]iven the unrelenting pace of technological progress, refusing to partake in the advantages offered by AI could place a large group of people at a serious disadvantage. As AI is reported and represented more and more in popular culture and in the media,

it could contribute to a deeply divided society, split between those who believe in (and consequently benefit from) AI and those who reject it.

From this perspective, mistrust is an obstacle, a speed-bump on the highway of inevitable progress, and skeptical users are pitted in opposition to the evocation of a greater public good. Not only is their skepticism detrimental to the latter; it will also be their own fault if they end up as the losers of the technosocial paradigm shift that AI is bringing about.

Such antagonistic positions, however, are less common in the discourse we reviewed, where reference to a decisively *symbiotic* relationship between AI systems, regulators, developers, and users is dominant. Rana el Kaliouby, CEO of an AI development firm specializing in human emotion detection, for example, writes on the World Economic Forum's website that "we're forging a new kind of partnership with technology. And with that partnership comes a new social contract: one that's built on mutual trust, empathy and ethics" (el Kaliouby 2019). This "new social contract", we hold, represents a misleading premise of equality between the actors that are designing and implementing AI, and the people who use it. The latter have usually very little knowledge about the system and its functioning; the system, by contrast, constantly gains information about the people—a problem that Zuboff (2019) in a similar context has identified as the decline of "reciprocity" (499–504).

Ideals of close partnership and symbiosis are further evoked when intergovernmental organizations recommend inclusive multi-stakeholder collaborations (UNESCO 2018) or prominent academics envision interdisciplinary collaborations across the public, private, and non-profit sectors driving the development of AI for social good (Tomašev et al. 2020). Common to all these is a tendency to obscure the power relations that underpin AI systems by conflating efforts for *building and managing* trust with the status of *deserving* trust.

The consequent outcome of such rhetorical maneuvers is the fostering of *blind trust* among the public, which is desired without attention to either the ways in which trust might be built, the reasons that it might not be merited, or what might go wrong. Indeed, this inherent opacity might also be the reason why, for all the emphasis on ethical AI in this discourse, attention to the potential harm it can cause is remarkably muted. While there are numerous examples of discriminatory and biased outcomes of AI use in the public sector, reinforcing socio-economic disparities and inequalities of justice (e.g.: Eubanks 2018; Grace 2019; O'Neil 2016; Park and Humphry 2019), the ways in which harm was caused are often unclear, because it remains an open question to what extent "algorithmic predictions influence human decision-makers" [Eubanks (2018), quoted Balaram et al. (2018, 10)].

4.2 Trust is a resource

Closely aligned with notions of blind trust is the rendering of *trust into a resource* and raw material—a process most readily apparent in the AI strategies of national governments, particularly of countries with high levels of trust. The Nordic Council of Ministers (2017), for example, states proudly that "The Nordic region is regarded as a world leader when it comes to social trust among its population", adding that public trust in government is "perhaps the most important resource in the Nordic societies" (10). The report pushes the issue even further by stating that "trust can be regarded as a *type of gold* for the Nordic countries" (Nordic Council of Ministers 2017, p. 7 our emphasis). In this way, trust, previously presumed to be an outcome and reward for virtuous governance, becomes commodified and hollowed out; framed as a resource to be manufactured, grown and managed, acquired and supplied to then be extracted and refined in the service of other aims.

True to this goldrush spirit, the Norwegian national strategy for AI (Norway 2020) frames its high levels of public trust as part of its national brand (cf., 2) and states its intention to lead "the way in developing human-friendly and trustworthy artificial intelligence" that, it hopes, "may prove a key advantage in today's global competition" (2). The Danish national strategy joins in: "[A]lmost all Danes use the internet on a daily basis, and there is a high degree of mutual trust and confidence. This means that Denmark is adaptable, and there is a good basis for implementing artificial intelligence" (Denmark 2019, p. 16).

However, also nations blessed with less of the 'Nordic gold' have come to understand trust as a resource. While Luxemburg (2019) intends to create "an innovative and trusted regulatory environment in order to attract data-driven and data-centric services and businesses" (13), the German national strategy (Germany 2018) articulates the hope that "the high level of data protection and privacy standards achieved in the EU build citizens' trust in new AI technologies and can therefore give German and European companies a competitive advantage internationally" (16). To briefly restate our core argument: While we deem achievements such as *high levels of data protection* as virtuous and welcome, the envisioned use of people's trust as a competitive advantage presents a slippery slope.

Just how much a social-engineering attitude towards trust as a resource has suffused the discourse on AI can be gathered from neologisms such as "undertrust" and "overtrust" (e.g., US National Science and Technology Council 2019), which imply that trust can be measured and weighed like the ingredients for a cake. This is disingenuous at best, in

a discourse that ambiguates between notions of trust as an *input* and an *output* of AI systems. In actuality, trust is far more entangled with other socio-economic and governance processes and outcomes (Wirtz et al. 2019) than is suggested by such neologisms. Ignoring those entanglements risks not only governance shortcomings, but also the normalization of particular agendas and interests that are served in the implementation of AI systems (Cath 2018).

4.3 Trust is an instrument

The notion of trust's commodification becomes especially salient if one assumes that trust has no inherent value in itself. Rather, the identification of *technologically* robust AI with *ethically* robust AI that we find in significant parts of the discourse is characterized by a strong instrumental undercurrent.

In these instances, trust is described as "essential" to "reap the full potential public and economic benefits from new technologies" (Mulgan 2016, p. 1), or to "fulfill the promise and value that AI can bring in sectors such as retail, finance, health care, and more" (Larsson et al. 2019, p. v). "The digital economy," states the Industrial Strategy of the UK's Department for Business (United Kingdom 2018), "relies on trust to work effectively" (159). And the strategy, "AI 4 Belgium", agrees: "After all, public trust is the cornerstone of any AI and data strategy. When it is lacking, innovation is off the table" (AI4Belgium Coalition 2019, p. 8).

Myers and Nejkov's (2020) describe this trend uncritically, noting that "ensuring trust has emerged as a precondition to realizing the social, commercial, and public benefits of implementing AI technologies" (1). However, while this statement is almost certainly true, it is indicative of a worrying trend which turns the public's trust into an instrument for unlocking other benefits: ensuring uptake and legitimacy of AI as well as its societal and market potentials. Indeed, this instrumental rationale makes analogies with "Nordic gold" and other commodities fall short. Rather than being a commodity in itself, trust in AI is desirable, because it catalyzes and enables the mining effort. It is not the final product, but—to present one last metaphor applied to it—a "lubricant" (Andreasson and Stende 2019, p. 28).

5 Conclusion

Our critical analysis of the discourse on AI and society suggests that there is an increasing trend towards the discursive commodification of trust. We find an overwhelming emphasis on the conditions for developing AI systems and unlocking markets, which conflates the interests and agendas of AI users, experts, and regulators, with those of technology vendors, venture capitalists, and financial backers, whose interests in advancing AI might be diametrically opposed. More critically, this emphasis has the effect of obfuscating the conditions under which trust between individuals and AI systems is developed and earned.

The insistence that governments focus on building public trust in AI systems obscures the questions about what might be required to actually build trustworthy systems. When trust is commodified in this discourse, there is no space to ask whether macro- or micro-level approaches are more appropriate, whether trust is justified, or who or what precisely users are asked to trust in. This might be dismissed as rhetorical emphasis; we hold, by contrast, that it is of critical importance, because the policy and scholarly discourse on AI and society is actively staking out the policy options for AI governance, and those governance structures will have far reaching and profound implications for society (Cath 2018; Cussins Newman 2020; Mulderrig et al. 2019). When appeals to build trust fail to acknowledge the intricacies of how trust gets built, they make the notion evermore elusive for both the designers of AI systems and the users.

Countering this discourse's gravitation toward blind, commodified, and instrumental trust, we would encourage experts, politicians, and stakeholders to instead focus on the work and the investments that indeed remain preconditional of the public's trust. Though not prominent in the discourse, there is solid research on which to base such an effort. Trustworthy AI may well require programming and design that is technologically robust (Bellamy et al. 2019; Etzioni and Etzioni 2016; Harrison et al. 2019; Kroll et al. 2017; Liao and Muller 2019; Sokol et al. 2020; Veale et al. 2018), processes and policy frameworks that protect citizens in the slightest case of doubt (Balaram et al. 2018; Brundage et al. 2020; Dignum 2019; Janssen and Kuk 2016; Kemper and Kolkman 2019; Kolkman 2020; Lee 2018; Mulgan 2016; Reisman et al. 2018; Vassilakopoulou 2020), or public-sector routines and protocols that allow for AI and human-service offers to run side-by-side for the foreseeable future, questioning, and learning from each other (Berscheid and Roewer-Despres 2019; Janssen et al. 2020; Katell et al. 2020; Rahwan 2017; Vestby and Vestby 2019; Yeung and Lodge 2019).

These might not be necessary mechanisms for building trust in AI, or the most appropriate or effective. But until the scholarly and policy discourse assessed here becomes actively engaged in an exploration of which mechanisms are the right mechanisms for building trust, the discourse on trust in AI is a shell game, baiting with market value and technological glitz, ignoring the hard questions about why trust matters. With the current discourse, governments and businesses are in no position to approach citizens and customers with a plea for their trust in shiny new systems. As Dignum (2019) writes: "Ensuring ethically aligned AI systems requires more than designing systems whose result can be trusted. It is about the way we design them, why we

design them, and who is involved in designing them" (v). We would add that it is also about fostering a discourse that actively acknowledges and engages with how this should be done.

Funding Open access funding provided by University of Oslo (incl Oslo University Hospital). This research was kindly supported by the Screen Cultures Initiative at the Faculty of Humanities, University of Oslo.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

References

AI4Belgium Coalition (2019) AI 4 Belgium. https://www.ai4belgium.be/wp-content/uploads/2019/04/report_en.pdf

Al-Mushayt OS (2019) Automating E-government services with artificial intelligence. IEEE Access 7:146821–146829. https://doi.org/10.1109/ACCESS.2019.2946204

Anderson J, Rainie L, Luchsinger A (2018) Artificial intelligence and the future of humans. https://www.pewinternet.org/2018/12/10/artificial-intelligence-and-the-future-of-humans/

Andreasson U, Stende T (2019) Nordic municipalities' work with artificial intelligence. Nord Munic Work Artif Intell. https://doi.org/10.6027/no2019-062

Arvanitakis J (2007) The commodification and re-claiming of trust: does anyone care about anything but the price of oil? Int J Interdiscip Soc Sci 2(3):41–50. https://doi.org/10.18848/1833-1882/CGP/v02i03/52313

Bachmann R, Inkpen AC (2011) Understanding institutional-based trust building processes in inter-organizational relationships. Organ Stud 32(2):281–301. https://doi.org/10.1177/0170840610397477

Balaram B, Greenham T, Leonard J (2018) Engaging citizens in the ethical use of AI for automated decision-making. https://www.thersa.org/globalassets/pdfs/reports/rsa_artificial-intelligence---real-public-engagement.pdf

Bellamy RKE, Mojsilovic A, Nagar S, Ramamurthy KN, Richards J, Saha D, Sattigeri P, Singh M, Varshney KR, Zhang Y, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S (2019) AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev. https://doi.org/10.1147/JRD.2019.2942287

Berscheid J, Roewer-Despres F (2019) Beyond transparency. AI Matters 5(2):13–22. https://doi.org/10.1145/3340470.3340476

Berzoff J (2011) Psychosocial Ego development: the theory of Erik Erikson. In: Berzoff J, Melano L, Hertz P (eds) Nside out and outside. Psychodynamic clinical theory and psychopathology in contemporary multicultural contexts, 3rd edn. Rowman & Littlefield, Lanham, pp 97–117

Bodó B (2021) Mediated trust: a theoretical framework to address the trustworthiness of technological trust mediators. New Media Soc 23(9):2668–2690. https://doi.org/10.1177/1461444820939922

Bourne C (2019) AI cheerleaders: public relations, neoliberalism and artificial intelligence. Public Relat Inq 8(2):109–125. https://doi.org/10.1177/2046147X19835250

Bourne C, Edwards L (2012) Producing trust, knowledge and expertise in financial markets: the global hedge fund industry "re-presents" itself. Cult Organ 18(2):107–122. https://doi.org/10.1080/14759551.2011.636614

Breeze R (2011) Critical discourse analysis and its critics. Pragmatics 21(4):493–525. https://doi.org/10.1075/prag.21.4.01bre

Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, Khlaaf H, Yang J, Toner H, Fong R, Maharaj T, Koh PW, Hooker S, Leung J, Trask A, Bluemke E, Lebensold J, O'Keefe C, Koren M, Anderljung M (2020) Toward trustworthy AI development: mechanisms for supporting verifiable claims. http://arxiv.org/abs/2004.07213

Candlin CN, Chrichton J (2013) From ontology to methodology: exploring the discursive landscape of trust. In: Candlin CN, Chrichton J (eds) Discourses of trust. Palgrave Macmillan, London, pp 1–20

Candlin CN, Crichton J (2013) Discourses of trust. In: Candlin CN, Crichton J (eds) Discourses of trust. Palmgrave Macmillan, London. https://doi.org/10.1007/978-1-137-29556-9

Cardullo P, Kitchin R (2019) Being a 'citizen' in the smart city: up and down the scaffold of smart citizen participation in Dublin, Ireland. GeoJournal 84(1):1–13. https://doi.org/10.1007/s10708-018-9845-8

Carolan MS, Bell MM (2003) In truth we trust: discourse, phenomenology, and the social relations of knowledge in an environmental dispute. Environ Values 12(2): 225–245. https://www.jstor.org/stable/30301940%0AJSTOR

Cath C (2018) Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philos Trans R Soc Math Phys Eng Sci 376(2133):20180080. https://doi.org/10.1098/rsta.2018.0080

Chakravorti B, Shankar R (2017) Digital planet 2017: how competitiveness and trust in digital economies vary across the world. The Fletcher School, Tufts University, Medford

Cukier W, Ngwenyama O, Bauer R, Middleton C (2009) A critical analysis of media discourse on information technology: preliminary results of a proposed method for critical discourse analysis. Inf Syst J 19:175–196. https://doi.org/10.1111/j.1365-2575.2008.00296.x

Cussins Newman J (2020) Decision points in AI governance (CLTC White Paper Series)

Dasgupta P (1989) Tust as commodity. In: Gambetta D (ed) Trust: making and breaking cooperative relations. Basil Blackwell, Hoboken, pp 49–72. https://doi.org/10.5860/choice.26-3664

Denmark (2019) National strategy for artificial intelligence of denmark (Issue March)

Dignum V (2019) Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer, Switzerland

Doteveryone (2018) People, power and technology: the 2020 digital attitudes report

El Kaliouby R (2019) How do we build trust between humans and AI? World Economic Forum. https://www.weforum.org/agenda/2019/08/can-ai-develop-an-empathetic-bond-with-humanity/

Erikson EH (1963) Childhood and society, 2nd edn. Norton, New York
Etzioni A, Etzioni O (2016) Keeping AI legal. Vanderbilt J Entertain Technol Law 19(1):133–146
Eubanks V (2018) Automating inequality: how high-tech tools profile, police, and punish the poor. St. Martin's Press, New York
European Commission (2020) White paper on artificial intelligence—a European approach to excellence and trust (COM(2020) 65 final). https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
European Strategy for Data (2020) A European strategy for data
Foucault M (1972) The Archaeology of Knowledge & the Discourse on Language. Pantheon Books, New York
Fairclough N (1995) Critical discourse analyisis – The critical study of language. Longman, London
Fairclough N (2010) Critical discourse analysis the critical study of language. Critical discourse analysis the critical study of language, 2nd edn. Rooutledge, Oxfordshire. https://doi.org/10.4324/9781315834368
Freuler JO, Iglesias C (2018) Algorithms and artificial intelligence in latin america: a study of implementation by governments in argentina and uruguay. https://webfoundation.org/research/how-are-governments-in-latin-america-using-artificial-intelligence/
Germany (2018) Artificial intelligence strategy
Giddens A (1990) The consequences of modernity. Polity Press, Cambridge
Grace J (2019) Machine learning technologies and their inherent human rights issues in criminal justice contexts Mr. Jamie Grace, Senior Lecturer in Law, Sheffield Hallam University, UK 1. PP. 1–19
Granovetter M (1985) Economic action and social structure: the problem of embeddedness. Am J Sociol 91:481–510. https://doi.org/10.4324/9780429494338
Hamm JA, Lee J, Trinkner R, Wingrove T, Leben S, Breuer C (2016) On the cross-domain scholarship of trust in the institutional context. Interdiscip Perspect Trust towards Theor Methodol Integr. https://doi.org/10.1007/978-3-319-22261-5_8
Hardin R (2002) Trust and trustworthiness. In: The Routledge handbook of trust and philosophy. Russel Sage Foundation, New York. https://doi.org/10.4324/9781315542294-2
Harrison TM, DePaula N, Luna-Reyes LF, Najafabadi MM, Pardo TA, Palmer JM (2019) The data firehose and AI in government: why data management is a key to value and ethics. ACM Int Conf Proc Ser. https://doi.org/10.1145/3325112.3325245
Henriques J, Hollway W, Urwin C, Venn C, Walkerdine V (1984) Changing the Subject. Psychology, Social Regulation and Subjectivity. Routledge, London
IBM (2018) IBM'S principles for data trust and transparency. IBM. Com. https://www.ibm.com/blogs/policy/trust-principles/
Janssen M, Kuk G (2016) The challenges and limits of big data algorithms in technocratic governance. Gov Inf Q 33(3):371–377. https://doi.org/10.1016/j.giq.2016.08.011
Janssen M, Brous P, Estevez E, Barbosa LS, Janowski T (2020) Data governance: organizing data for trustworthy artificial intelligence. Gov Inf Q 37(3):101493. https://doi.org/10.1016/j.giq.2020.101493
Katell M, Young M, Dailey D, Herman B, Guetler V, Tam A, Binz C, Raz D, Krafft PM (2020) Toward situated interventions for algorithmic equity: lessons from the field. FAT* 2020—Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 45–55. https://doi.org/10.1145/3351095.3372874
Kemper J, Kolkman D (2019) Transparent to whom? No algorithmic accountability without a critical audience. Inf Commun Soc 22(14):2081–2096. https://doi.org/10.1080/1369118X.2018.1477967
Kolkman D (2020) The (in)credibility of algorithmic models to non-experts. Inf Commun Soc. https://doi.org/10.1080/1369118X.2020.1761860
Kroll JA, Barocas S, Felten EW, Reidenberg JR, Robinson DG, Yu H (2017) Accountable algorithms. Univ Pa Law Rev 165(3):633–706. https://doi.org/10.1017/CBO9781107415324.004
Kuziemski M, Misuraca G (2020) AI governance in the public sector: three tales from the frontiers of automated decision-making in democratic settings. Telecommun Policy 44(6):101976. https://doi.org/10.1016/j.telpol.2020.101976
Larsson S, Anneroth M, Felländer-Tsai L, Institutet K, Heintz F, Ångström RC (2019) Sustainable AI: an inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence. http://www.aisustainability.org/wp-content/uploads/2019/04/SUSTAINABLE-AI.pdf
Lee MK (2018) Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. Big Data Soc 5(1):1–16. https://doi.org/10.1177/2053951718756684
Liao QV, Muller M (2019) Enabling value sensitive AI systems through participatory design fictions. Preprint
Luhmann N (2017) Tust and power. Polity Press, New York
Luxembourg (2019) Artificial intelligence: a strategic vision for Luxembourg. https://doi.org/10.1056/NEJM198006263022618
Matheny M, Israni ST, Ahmed M (Eds.) (2019) Artificial intelligence in health care: the hope, the hype, the promise, the peril. National Academy of Medicine. https://doi.org/10.1017/CBO9781107415324.004
Mazey NCHL, Wingreen SC (2017) Perceptions of trust in bionano sensors: is it against our better judgement? An investigation of generalised expectancies and the emerging technology trust paradox. Int J Distrib Sens Netw. https://doi.org/10.1177/1550147717717388
Mehr H (2017) Artificial intelligence for citizen services and government. https://ash.harvard.edu/files/ash/files/artificial_intelligence_for_citizen_services.pdf
Metzinger T (2019) EU guidelines: ethics washing made in Europe. Tag Spiegel. https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html
Mikhaylov SJ, Esteve M, Campion A (2018) Artificial intelligence for the public sector: opportunities and challenges of cross-sector collaboration. Philoso Trans R Soc. https://doi.org/10.1098/rsta.2017.0357
Misuraca G, Van Noordt C (2020) AI watch: artificial intelligence in public services—overview of the use and impact of AI in public services in the EU (EUR 30255). Publ off Eur Union. https://doi.org/10.2760/039619
Mulderrig J, Montesano N, Farrelly M (2019) Introducing critical policy discourse analysis. In: Montessori NM, Farrelly M, Mulderrig J (Eds.). Critical policy discourse analysis. Edward Elgar Publishing, Northampton, pp. 1–22. https://www.elgaronline.com/view/edcoll/9781788974950/9781788974950.xml
Mulgan G (2016) A machine intelligence commission for the UK: how to grow informed public trust and maximise the positive impact of smart machines. https://media.nesta.org.uk/documents/a_machine_intelligence_commission_for_the_uk_-_geoff_mulgan.pdf
Myers G, Nejkov K (2020) Developing artificial intelligence sustainably: toward a practical code of conduct for disruptive technologies. In https://www.Ifc.Org/Wps/Wcm/Connect/Publications_Ext_Content/Ifc_External_Publication_Site/Publications_Listing_Page/Emcompass-Note-80-Tocc (No. 80; EM Compass Notes)
Nordic Council of Ministers (2017) Trust—the Nordic Gold
Norway (2020) Nasjonal strategi for kunstig intelligens
O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Crown Random House, New York
OECD (2019) OECD principles on artificial intelligence. In: Organisation for economic co-operation and development. organisation for

economic co-operation and development. http://www.oecd.org/going-digital/ai/principles/
Oravec JA (2019) Artificial Intelligence, automation, and social welfare: some ethical and historical perspectives on technological overstatement and hyperbole. Ethics Soc Welf 13(1):18–32. https://doi.org/10.1080/17496535.2018.1512142
Orr W, Davis JL (2020) Attributions of ethical responsibility by artificial intelligence practitioners. Inf Commun Soc. https://doi.org/10.1080/1369118X.2020.1713842
Park S, Humphry J (2019) Exclusion by design: intersections of social, digital and data exclusion. Inf Commun Soc 22(7):934–953. https://doi.org/10.1080/1369118X.2019.1606266
Pennington R, Dixon Wilcox H, Grover V (2003) The role of system trust in business-to-consumer transactions. J Manag Inf Syst 20(3):197–226. https://doi.org/10.1080/07421222.2003.11045777
Polonski S (2018) AI trust and AI fears: a media debate that could divide society. Medium. https://medium.com/@drpolonski/ai-trust-and-ai-fears-a-media-debate-that-could-divide-society-52e16a74c979#id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6ImUxOTdiZjJlODdiZDE5MDU1NzVmOWI2ZTVlYjQyNmVkYTVkNTc0ZTMiLCJ0eXAiOiJKV1QifQ.eyJpc3MiOiJodHRwczovL2FjY291bnRzLmdv
Rahwan I (2017) Society-in-the-loop: programming the algorithmic social contract. Ethics Inf Technol 20(1):5–14. https://doi.org/10.1007/s10676-017-9430-8
Reisman D, Schultz J, Crawford K, Whittaker M (2018) Algorithmic impact assessments: a practical framework for public agency accountability. https://ainowinstitute.org/aiareport2018.pdf
Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you? Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, August, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778
Shapiro SP (1987) The social control of impersonal trust. Am J Sociol 93(3):623–658. https://doi.org/10.1086/228791
Sokol K, Hepburn A, Poyiadzi R, Santos-rodriguez R, Flach P (2020) FAT forensics: a python toolbox for implementing and deploying fairness, accountability and transparency algorithms in predictive systems. J Open Sour Softw 5:10–13. https://doi.org/10.21105/joss.01904
Steedman R, Kennedy H, Jones R (2020) Complex ecologies of trust in data practices and data-driven systems. Inf Commun Soc. https://doi.org/10.1080/1369118X.2020.1748090
Stoddart MCJ (2007) Ideology, hegemony, discourse : a critical review of theories of knowledge and power. Soc Thought Res 28:191–225
Tomašev N, Cornebise J, Hutter F, Mohamed S, Picciariello A, Connelly B, Belgrave DCM, Ezer D, van der Haert FC, Mugisha F, Abila G, Arai H, Almiraat H, Proskurnia J, Snyder K, Otake-Matsuura M, Othman M, Glasmachers T, de Wever W, Clopath C (2020) AI for social good: unlocking the opportunity for positive impact. Nat Commun 11(1):2468. https://doi.org/10.1038/s41467-020-15871-z
Toreini E, Aitken M, Coopamootoo KPL, Elliott K, Zelaya VG, Missier P, Ng M, van Moorsel A (2020) Technologies for trustworthy machine learning: a survey in a socio-technical context. ArXiv:2007.08911v1. http://arxiv.org/abs/2007.08911
Trust: making and breaking cooperative relations (1989). In: D Gambetta (Ed.). Choice reviews online, Vol. 26, Issue 07. Basil Blackwell. https://doi.org/10.5860/choice.26-3664
UNESCO (2018) Steering AI and advanced ICTs for knowledge societies human rights implications. https://en.unesco.org/system/files/unesco-steering_ai_for_knowledge_societies.pdf
United Kingdom. (2018). Industrial strategy artificial intelligence sector deal. https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal
US National Science and Technology Council (2019) The national artificial intelligence research and development strategic plan: 2019 update. https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf
Vassilakopoulou P (2020) Sociotechnical approach for accountability by design in AI systems
Veale M, Van Kleek M, Binns R (2018) Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. Conference on human factors in computing systems—proceedings, pp. 1–14. https://doi.org/10.1145/3173574.3174014
Vestby A, Vestby J (2019) Machine learning and the police: asking the right questions. Polic J Policy Pract. https://doi.org/10.1093/police/paz035
Villani C (2018) For a meaningful artificial intelligence: towards a French and European strategy
Wirtz BW, Weyerer JC, Geyer C (2019) Artificial Intelligence and the public sector—applications and challenges. Int J Public Adm 42(7):596–615. https://doi.org/10.1080/01900692.2018.1498103
World Economic Forum (2017) The global risks report 2017 12th Edn. Summary for policymakers. https://doi.org/10.1017/CBO9781107415324.004
Yeung K, Lodge M (2019) Algorithmic regulation. In: Yeung K, Lodge M (eds) Algorithmic regulation. Oxford University Press, Oxford, pp 1–18. https://doi.org/10.1093/oso/9780198838494.003.0001
Zuboff S (2019) The age of surveillance capitalism. Profile
Zucker LG (1986) Production of trust: institutional sources of economic structure, 1840–1920. Res Org Behavi 8:53–111