Original Research Article

BIG DATA & SOCIETY

Big Data & Society
April–June: 1–14
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517241249430
journals.sagepub.com/home/bds
Sage

The trustification of AI. Disclosing the bridging pillars that tie trust and AI together

Jascha Bareis[1,2]

Abstract

Trustworthy artificial intelligence (TAI) is trending high on the political agenda. However, what is actually implied when talking about TAI, and why it is so difficult to achieve, remains insufficiently understood by both academic discourse and current AI policy frameworks. This paper offers an analytical scheme with four different dimensions that constitute TAI: a) A user perspective of AI as a quasi-other; b) AI's embedding in a network of actors from programmers to platform gatekeepers; c) The regulatory role of governance in bridging trust insecurities and deciding on AI value trade-offs; and d) The role of narratives and rhetoric in mediating AI and its conflictual governance processes. It is through the analytical scheme that overlooked aspects and missed regulatory demands around TAI are revealed and can be tackled. Conceptually, this work is situated in disciplinary transgression, dictated by the complexity of the phenomenon of TAI. The paper borrows from multiple inspirations such as phenomenology to reveal AI as a quasi-other we (dis-)trust; Science & Technology Studies (STS) to deconstruct AI's social and rhetorical embedding; as well as political science for pinpointing hegemonial conflicts within regulatory bargaining.

Keywords

Trustworthy AI, technology governance, conflict theory of state, ethics of AI, public interest AI, science and technology studies

Introduction

Trustworthy artificial intelligence (TAI) is trending high on the political agenda. The advancement of artificial intelligence (AI) technology has been endowed with massive investments and great hopes by governments around the world to solve pressing problems in our societies. However, past incidents related to AI have provoked attention and outcry in media and led to hesitation to continue down the path of AI enthusiasm unquestioningly. AI can be misused to manipulate political opinion with deep fakes (van Huijstee et al., 2021). The COMPAS recidivism risk assessment tool used in the US judiciary paradigmatically shows how incidents of bias and discrimination in data processing can aggravate racism and inequality in criminal prosecution (Angwin et al., 2016). Or, while crucial infrastructure becomes ever more automated with AI, issues of safety, robustness and network vulnerability arise from failing systems (McMillan and Varga, 2022). These are only indicative examples that show some salient problems with AI systems.

Such publicly discussed incidents pose a great threat to building and maintaining trust in AI systems and in the institutions that provide these systems and protect users. Faced with these individual and systemic impacts of AI on our societies, regulators are on the spot to carefully weigh the potentials and risks and develop effective policy. As a result, nation states have addressed the urgency of developing policies that address users' ethical concerns while harvesting the economic and efficiency benefits of AI in strategy and position papers (Radu, 2021). However, while there is a growing emphasis on the trust dimension in AI governance in these papers, the pairing of trust and AI is far from intuitive. It invokes first and foremost an unorthodox relationship: It marries a widely *technically* employed term, AI, with a *social* one, trust. How

[1] Researcher, Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2] Associate researcher, Humboldt Institute for Internet and Society, Berlin, Germany

Corresponding author:
Jascha Bareis, Researcher, Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology, Karlsruhe, Germany.
Email: jascha.bareis@kit.edu

to bridge this technical to social domain is not so obvious and straightforwardly answered (see section *Pairing trust and AI – a conceptual challenge*).

Why should policy makers and researchers care about trust in the governance of a multifaceted technology like AI? First, to understand the general value of trust for technology governance, it is helpful to recognise that distrust in particular can be very costly for society (Hardin, 2002; Warren, 1999). In general, trust relationships are characterised by a state of uncertainty and risk (Luhmann, 1988; Misztal, 1996). If users had perfect knowledge and control over their technological environment, notions of trust in technology would be redundant. People are willing to give up control if they can be sure that their peers will not act against their interests (Coleman, 1986). Put simply, if one trusts and gives up control, one can save and/or redirect resources. Distrust not only does the opposite, it can be lastingly damaging as it ruins reputations and leads to a great loss of social and economic capital (North, 1990). When distrust spreads and becomes endemic, everyone infected loses. AI scandals illustrate this phenomenon. In the worst cases, users feel betrayed, AI applications are rejected, providers are boycotted, money is burned, and governments' regulatory capacity is questioned. But this also implies that distrust is not always negative. Citizens signalling distrust can also represent a healthy watchdog mechanism for checks and balances, for example by flagging misplaced or badly executed AI systems, regulatory capture or empty rhetoric[1].

Second, and very concretely, TAI plays a pivotal role in the current regulatory debate, as it spearheads regulatory frameworks such as the European AI Act (AIA). Unfortunately, the regulatory approach to trust so far has been rather vague and confusing, lacking definitions and a deeper understanding of trust (section *Trustworthy AI in the current landscape: From ethical values to regulatory frameworks*).

Third, the current ethical and regulatory debate on TAI is very much fixated on a technical understanding of AI (section *Opening technical AI to social dimensions of trust*) and its debugging of harmful effects, such as providing computational methods in inspecting models and providing interpretability (see discussion by Páez, 2019; Zednik, 2019; and von Eschenbach, 2021), or de-biasing, discussing trade-offs between algorithmic efficiency and different variations of fairness (Kleinberg et al., 2016; Wong, 2020). This technical debate has its merit, but it lost track of the actual social preconditions that tie trust and AI together.

Therefore, this paper responds to current governance initiatives and ethical discussions that invoke trust as an important variable in AI regulation. To be clear from the start: The main aim of this paper is *not* to assess whether AI is trustworthy or not, but to give an account of the dimensions that need to be considered in order to *be able* to assess it. Hence, first of all, this paper takes a step back and revisits the concepts of trust and AI. Given the contested relationship between the two phenomena, what are the epistemic dimensions that tie trust and AI together? To answer this research question I forward and execute an analytical scheme based on four pillars: a) AI as a quasi-other; b) AI's embedding in a network of actors from programmers to platform gatekeepers; c) the regulatory role of governance in bridging trust insecurities and deciding on AI value trade-offs; and d) the role of narratives and rhetoric in mediating AI its conflictual governance processes. It is through this systematization that overlooked aspects and missed regulatory demands around TAI are revealed and can be addressed (see *Concluding remarks*).

This work can be understood as a follow-up on comprehensive systematization works on trust in information and communication technologies (ICT), such as in e-commerce (McKnight et al., 2002), in information systems (Söllner et al., 2016), or in broader readings of technology (Botsman, 2017). However, the complexity of the AI phenomenon requires both different analytical and disciplinary approaches than the ones targeting ICT systems. Therefore, this work borrows from multiple academic viewpoints and concepts. Among other, I refer to phenomenology in order to reveal AI as a quasi-other that we (dis)trust; Science & Technology Studies (STS) to deconstruct the social and rhetorical embedding of AI; and political science to identify hegemonic conflicts in regulatory bargaining. This, admittedly, wide approach is less a scholarly preference but owed to the complexity of the AI phenomenon itself. With disciplinary blinkers, one would miss the constitutive bridging pillars that connect trust with AI. In my approach, I adhere to the agenda of critical algorithmic studies, which is "essentially, founded in a disciplinary transgression" (Seaver, 2017: 2).

Trustworthy AI in the current landscape: From ethical values to regulatory frameworks

Ethical principles. Recently, there has been a rich landscape of TAI work emerging in both academic debate and governance proposals. The publication of ethical guidelines has reached a scale that is hard to keep track of[2]. High-level principles are published by political bodies and by big Tech companies that aim to ensure a socially desirable implementation of AI, linking ethical values to notions of trustworthiness (EU High-Level Expert Group on AI, 2019; European Commission, 2020a; OECD, 2019). The most dominant approach towards TAI is embedded in the field of ethics. Here, trust is operationalised as a resulting phenomenon that emerges from following a checklist of ethical requirements that need to be 'handled' or 'taken care of'. In this strikingly instrumental understanding of trust, ethicists list values, such as transparency, privacy, accountability, fairness or robustness as fundamental requirements. Kaur et al. (2023) and Reinhardt (2022) undertake great efforts in assembling all the literature of TAI that unites behind each of these single ethical values (see also Simion and Kelp, 2023).

The ethical discourse, even when condensed,[3] is descriptively rich but at the same time abundant and abstract, lacking clarity and consensus. Lists of axiomatic AI principles from the public and private sector levitate over the contested reality of society. It is implied that ethical values can be analytically 'isolated', thereby failing to point to the ambivalences and tensions arising between the values (Mittelstadt, 2019). Furthermore, the overall difficulty and reservation to operationalize normative principles and rights into quantitative and measurable scores for governance, while isolating them from their social surrounding and context (Hoffmann, 2019), has led some to bluntly conclude that the discourse of AI ethics is essentially "useless" (Munn, 2022). On a rather poetic remark, Reinhardt (2022) observes that the academic field of trust and AI has turned into "an intellectual land of plenty, a mythological or fictional place where everything is available at any time without conflicts" (741). In conclusion, ethical values may give guidance for better understanding the risks associated with AI but little can be deduced from the ethical discourse in better understanding the phenomenon of TAI.

Governing frameworks. This ethical discourse is flanked by the crafting of a global governance regime around AI. So far, this regime consists of an overlapping ensemble of private standards, normative principle-setting, concrete standardization efforts, as well as the creation of new legal frameworks that shall extend or replace existing (inter-)national legislation (Veale et al., 2023). Supranational bodies such as the OECD (2019) recommended some guiding (albeit again vague) principles for TAI which it would like to see promoted and implemented, taken up by the United Nations which published a more detailed interim Report on "Governing AI for Humanity" in late 2023.

Of all global players, the EU has unquestionably been most proactive in coming up with a coercive and unified framework for establishing TAI[4]. The AIA passed the European Parliament in March 2024 and will come into force by 2026 (European Commission, 2024). The EU commission had initiated the negotiation process in 2019 to develop a distinct European approach to "Excellence and Trust in Artificial Intelligence" (European Commission, 2020b). In the same year, the High-Level Expert Group on AI (HLEG) set the normative foundations for EUs understanding of TAI, forwarding some ethical principles derived from the EU fundamental rights framework (High-Level Expert Group, 2019). The 2020 European Commission White Paper, embedded in a public consultation process, similarly stressed: "As digital technology becomes an ever more central part of every aspect of people's lives, people should be able to trust it. Trustworthiness is also a prerequisite for its uptake" (European Commission, 2020b: 1), following up with a bold proposal of an "ecosystem of trust".

This AI EU ecosystem of trust builds on three pillars (5):

> "1. it should be lawful, complying with all applicable laws and regulations;
>
> 2. it should be ethical, ensuring adherence to ethical principles and values; and
>
> 3. it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm."

The final text of the AIA comprises these pillars combined with the economic argument of establishing a common AI market integration. The AIA clarifies that "[t]he purpose of this Regulation is to improve the functioning of the internal market and promoting the uptake of human centric and trustworthy artificial intelligence (…)" (European Commission, 2024: 93). In its regulatory paradigm, the AIA combines a principle-based framework of rights with a risk regulatory assessment of harms, while simultaneously aiming at an innovative and internationally competitive AI market (Krarup and Horst, 2023). In its final version the AIA proposal prescribes various instruments of risk regulation, organised around four risk categories, where each AI application is categorised before entering the market.

The notion of trust enters the picture with the classification of high-risk AI systems. They are handled through a self- and third-party conformity assessment (AIA, Article 43). Such assessment builds on the 2020 'self-assessment list for Trustworthy Artificial Intelligence' (ALTAI), which can be understood as a technical and ethical check list (European Commission, 2020a). If this self-assessment or third-party assessment will be enforced in a rigorous and effective way remains disputable, given the general contestability and interpretative vagueness of ethical values and the questionable willingness of profit seeking companies to curtail themselves with higher conformity obligations. Here, users will simply have to trust providers and third-parties. Interestingly, trust is rather featured as a European selling point in the AIA than really being defined. "The Act portrays this declaration of conformity with EU standards as a chief marker of "trustworthiness" (Paul, 2023:12). Thus, it is the *entire* EU conformity system that is branded as trustworthy, without any explanation of what is essentially meant by trust in the context of AI. It is striking that the entire EU regulatory framework lacks a *single* definition of trust. As a result, the presentation of TAI in EU documents appears slightly circular. In a nutshell: The EU AI regulation is trustworthy because AI is addressed by the EU.

The term TAI lacks semantic quality. As will be shown, this is problematic because regulation risks missing core dimensions of trust that are important for the governance of AI.

Pairing trust and AI – a conceptual challenge

Towards a sincere understanding of trust. Before delving into the different dimensions of trust in AI (section *Trust dimensions in AI*), the following section clarifies what to actually look for. Trust is not an axiomatic ethical value as the current ethical debate on AI might suggest. To refer to the introductory remarks, trust is a phenomenon that emerges in the social interaction of individuals and collectives characterised by risk and uncertainty. Conceptual and analytical debates on trust focus on the different reasons for entering into trust relationships and on the characteristics of the trust-giver, the trust-taker, and their relationship. Here, trust is generally understood as a social attitude, a normative, mostly emotional expectation towards an entity x and its performance (Hardin, 2002; McLeod, 2021). Trustworthiness, in turn, is a quality or characteristic of entity x and its performance that motivates to provide sufficient reason to justify the attitude of trust (Nickel, 2013). The commonly used analytical scheme to analyse trustworthiness is a three-place relationship: “B is trustworthy for person A with regard to the performance of x” (Nickel et al., 2010: 431). Applying this analytical scheme to the technological domain is neither intuitive nor unproblematic. The dominant approach to trustworthy technology relates to the factor of functionality, which is understood as reliability in performance. “Reliability is a characteristic of an item, expressed by the probability that the item will perform its required function under given conditions for a stated time interval” (Nickel et al., 2010: 433). It should be noted that the connotation of reliability is heavily influenced by an engineering and rational choice perspective that links the performance of technology to the risk of failure, for example, the risk of infrastructure collapsing.

However, many scholars argue that reducing trust to the notion of reliability does not do justice to the true nature of trust, raising the question of whether one should use the concept of trust at all in the context of technology. They link trust to a richer notion that requires some motivation, also known as ‘motive-based’ theories of trust. These scholars argue that trust must include motives of goodwill and notions of betrayal, thus emphasising emotional involvement (Baier, 1986; Jones, 1996). Others argue that there must be a moral dimension present, such as moral integrity or a person bound by a moral obligation, in order to speak of trust relationships (McLeod, 2002; Nickel, 2007). These broader conceptions of trust defend trust as an inherently interpersonal phenomenon. Trust is conceptualised as a uniquely human feature, capable of emotions, agency and moral intentions, rather than a phenomenon between objects or technology. The enthusiasm of some thinkers commenting on the pairing of trust and technology is rather reserved. Jones writes: “Trusting is not an attitude that we can adopt toward machinery. I can rely on my computer not to destroy important documents or on my old car to get me from A to B, but my old car is reliable rather than trustworthy. One can only trust things that have wills (…)” (Jones, 1996: 14; see also Ryan 2020 on AI). These reservations about simply transferring interpersonal trust to human-machine trust are instructive for the TAI debate. If one wants to pair trust and AI, one needs to look for features that characterise human-machine relationships beyond reliability.

Opening technical AI to social dimensions of trust. Finding these social and uncertainty realms acknowledges a broader understanding of AI. There is a plethora of definitions of AI coming from academia, corporations, tech gurus and policy papers. Certain features of AI are favoured in certain disciplines, reflecting the diversity of existing AI applications and research. This abundance of discourse has unfortunately led to much confusion around the term in both policy (Folberth et al., 2022) and in public discourse (Natale and Ballatore, 2020) (see also *Promoting trustworthy AI through narratives: mediating meaning & attention).*

From a technical perspective, AI applications aim to perform some ideal action or reasoning associated with mimicking human tasks and thinking (Krafft et al., 2020). Due to recent technical developments in data processing capabilities and the implementation of statistical learning theory, machine learning (ML) has become the state of the art in AI applications, alongside logic and knowledge-based approaches (Russell and Norvig, 2022). ML relies on great access to data to make robust predictions and to correct performance errors in iterative computational sequences. The technical focus of AI is also dominant in policy papers. For example, the AIA, Art. 3, defines AI as a “machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that (…) can influence physical or virtual environments”.

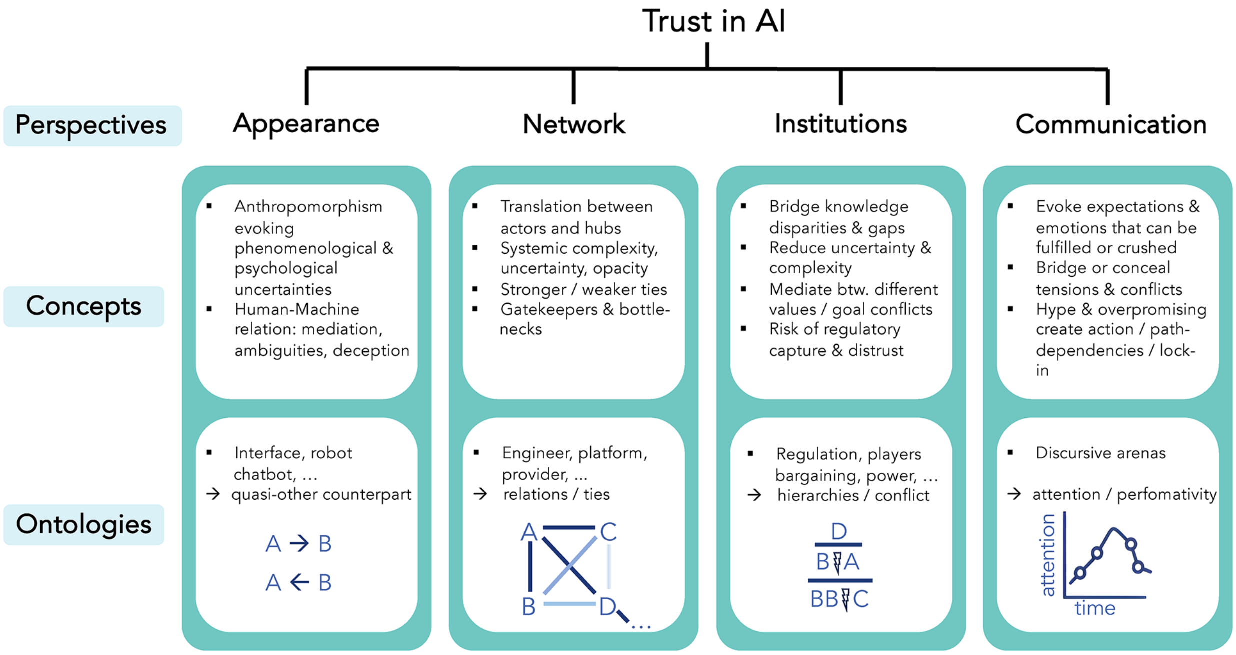
Surprisingly, *social* environments are not part of the AIA AI definition and that is problematic if one wants to understand the role of trust in the picture. While technical definitions may suggest delimitation and clarity, they fall short of a larger notion when it comes to encompassing AI’s relationship to trust. They fail to capture the distinct phenomena that AI applications produce, which arise not so much from algorithmic performativity but the meaning that is ascribed to it. I argue that AI is not only embedded in the social - but is *constituted* by it. The way AI is perceived and approached by users, embraced by institutions, praised by tech-gurus, and talked about in media points to a constant and complex dynamic between the actual technological developments and the potentials, fears and futures that are associated with it. It is exactly this constant *tension* between fact and fiction, hype and reality, scandal and breakthrough which is rendering AI so performative as a social phenomenon. I follow a reading that builds on an understanding of AI as situated and relational (Suchman, 2023; Suchman and Weber, 2016; Mackenzie, 2015), reworked and understood by different users and enmeshed in constellations of power. AI is hardly perceived and approached as a clearly articulated, delimited, and external 'thing', 'model' or 'tool' like some technical definitions suggest. Also, in their daily interaction users actually never see code, databases or backends of AI applications. Rather than approaching AI as a self-standing entity that can be generalised ('AI is x'), in this reading AI is woven and negotiated in the everyday realities of users and society, with its applications mediating human relationships, producing intimacies, social orders and knowledge authorities. It is exactly in this dynamic sphere that I will place the analysis of the following sections, as it is here that one can locate the constitutive bridging pillars that tie trust and AI together. The upcoming scheme (see Table 1) should be understood as an offer to policymakers and researchers when they invoke trust relationships with AI, doing justice to the complexity and fragility of the phenomenon. Building trust is challenging, but also rewarding. As outlined in the introduction, respecting the role that (dis)trust plays in the acceptance and rejection of technology is central to designing successful policies.

Trust dimensions in AI

Phenomenological appearance: trusting AI as a quasi-other. From its very beginnings - the foundation of modern AI in the 1950s - AI has been associated with the phenomenon of anthropomorphism: the attribution of human characteristics to objects, behaviours or features - in this case, machines (Salles et al., 2020). In 1966, the computer scientist Joseph Weizenbaum fed his chatbot ELIZA with the DOCTOR script, imitating a Rogerian psychotherapist. ELIZA was a very rudimentary chatbot, programmed to simply rephrase patients' answers as backfeed questions (Güzeldere and Franchi, 1995). Weizenbaum was struck when he observed that his chatbot elicited very emotional and intimate responses from his probands. What has since become known as the 'ELIZA effect' is a powerful demonstration of how humans can project emotions onto machines. The experiment shows that it is not so much the human-like capabilities of algorthmic decision making programs that trigger anthropomorphism (since ELIZA was a very simple software), but their combination with the vast field of human imagination. It is this combination of suggestive human characteristics of a machine with the power of human imagination that enables the emotional attachment to AI, whether it be social robots, assistive interfaces, or recent large-language-model chat bots like 'Chat-GPT' or 'Gemini'.

Recent academic discourses such as postphenomenology or robot-ethics have elaborated new epistemologies for technological mediation. They develop new concepts of human-machine interaction (Latour, 1994) and technology embodiment (Ihde, 2009; Suchman, 2007); or discuss whether robots appearing in our social world should be understood as moral agents with rights (Loh, 2019; Wallach and Allen, 2008). Without entering into the discussion of whether it being legitimate or helpful to call AI systems moral agents with wills, it is an empirical fact that they increasingly appear human and interact

Table 1. Four different trust dimensions that constitute TAI. Visualizing the metastructure of the upcoming analysis.[6]



Trust in AI

Perspectives	Appearance	Network	Institutions	Communication
Concepts	▪ Anthropomorphism evoking phenomenological & psychological uncertainties ▪ Human-Machine relation: mediation, ambiguities, deception	▪ Translation between actors and hubs ▪ Systemic complexity, uncertainty, opacity ▪ Stronger / weaker ties ▪ Gatekeepers & bottle-necks	▪ Bridge knowledge disparities & gaps ▪ Reduce uncertainty & complexity ▪ Mediate btw. different values / goal conflicts ▪ Risk of regulatory capture & distrust	▪ Evoke expectations & emotions that can be fulfilled or crushed ▪ Bridge or conceal tensions & conflicts ▪ Hype & overpromising create action / path-dependencies / lock-in
Ontologies	▪ Interface, robot chatbot, … → quasi-other counterpart	▪ Engineer, platform, provider, … → relations / ties	▪ Regulation, players bargaining, power, … → hierarchies / conflict	▪ Discursive arenas → attention / perfomativity

with us as "quasi-others" (Coeckelbergh, 2012: 75). The recent use of AI in the field of personal assistance technologies based on natural language processing, such as Apple's 'Siri' or Amazon's 'Alexa' (Silva de Barcelos et al., 2020), social robotics applied in the fields of care, elderly and sex services (Scheutz and Arnold, 2016; Sheridan, 2020), or the use of user interfaces at work (Bader and Kaiser, 2019) are very indicative in this regard.

The phenomenological perspective makes clear that AI systems, even if they only *simulate* human characteristics such as motivations, morals and emotions, can raise expectations of trust. When people interact intimately with AI systems, they embark on fragile social bonds and expose themselves to emotional attachments. In doing so, they are confronted with a core characteristic of trust: the loss of control. When I show intimate emotions, I expose myself vulnerable as I develop expectations that can trigger feelings of validation, resentment or even betrayal. For the motive-based theorists of trust mentioned above, this phenomenological perspective may be frustrating because it refers only to projections and simulations of social beings, but this does not make it any less attractive to many human interactants. Undoubtedly, societies are only at the beginning of an increasing conflation of the real and the virtual, as AI applications are implemented in all kinds of social spheres.

AI as a quasi-other appears not only in social robotics or interfaces, but also with synthetically generated content. The flooding of the internet with deep fakes or factually false content generated by large-language-models has become a major concern in politics. Here, the blurring is deliberate and systematically aimed at disinformation and manipulation of users and the public, threatening the free formation of opinion and the personal integrity of individuals (Chesney and Citron, 2019; van Huijstee et al., 2021). The weaponisation of suspicion and distrust has already sparked a deliberate military coup in Gabun in January 2019, where a (quite rudimentary) deep fake video of Gabun's President Ali Bongo appearing numb and motionless went viral amid public speculation about his health condition (Washington Post, 2020).

Conclusively, this section stresses that AI as an intersubjective, quasi-other is a pivotal analytical dimension for understanding the relationship between trust and AI. In the face of AI challenging and blurring reality, regulators are on the spot to intervene. So far, the EU AIA imposes transparency duties on the producer of synthetic content, requiring it to be labelled (Art. 52 III). Synthetically produced content will soon increase in quantity and quality and producers will be harder to identify or deliberately remain anonymous villains. Who will be responsible for identifying and proving what is fake or real in the digital world - and will it even be technically possible to distinguish between these states in the future? What content can users trust or must distrust? Current regulatory frameworks fail to address this gap. While the EU's Digital Service Act (DSA) (European Commission, 2022b) prescribes a "notice and take down action" procedure for digital platforms (Art. 14, 14 III, 19), it comes with a caveat. Platforms are not obliged to actively monitor any content and are exempt from liability for the distribution of illegal content as long as they are not aware of it. They wait to be notified by users to flag illegal and offensive content. This, of course, externalises corporate accountability and leaves considerable room for loopholes.

What current TAI governance discussion is missing completely, though, is a reflection of where to draw the line on the role(s) AI should take as quasi-others in very intimate spheres of society such as care, child education or sexuality. It is here where trust relationships are most fragile and people are most exposed and vulnerable. Individuals are already revealing their most intimate selves to AI applications and to much more rudimentary algorithmic systems (see ELIZA). The intrusion of AI into intimate spheres radically puts society's emotional and moral worldviews up for negotiation, as humans are lured out of their comfortable and taken for granted anthropocentric comfort zones. Which boundaries between humans and AI are still legitimate and to be trusted, which even need to be maintained? So far, policymakers have provided little guidance on these questions, and societies are navigating rather blindly into an increasingly blurring of the analogue and the digital, the authentic and the fake.

Trust the network. AI's social embedding and platformization. The relationship between AI and trust is not only demarcated by an intersubjective and apparent quasi-other. Many factors in a muted and hidden structural background play a key role in trust, embedding an AI application in a network of relationships between different actors. Among others: company leaders, designers, engineers, clickworkers, policy makers, users, and non-users. This extends the network of trust beyond the technological application. Von Eschenbach (2021) notes: "Trust with respect to technology (…) can only be understood in reference to the system as a whole, and each agent's trustworthiness will be judged relative to the differences in roles, interests, and expertise" (1619). The EU HLEG also stresses the importance of a systemic trust account: "Trustworthy AI (…) concerns not only the trustworthiness of the AI system itself, but requires a holistic and systemic approach, encompassing the trustworthiness of all actors and processes that are part of the system's socio-technical context throughout its entire life cycle" (2019: 5). In effect, the notion of trust is extended from AI as an application to a web of different actors involved in the chain of building and delivering a trustworthy AI system.

In addition to the concealed social and technical background processes inherent to the respective AI system, AI applications are embedded in different use contexts and domains. Today, societies are beginning to implement AI in all fields, whether it is work, health, entertainment, military or administration. AI systems act as sorting systems that decide who to hire or not (Laurim et al., 2021), mediate users' access to information through recommender systems on platforms (Gorwa et al., 2020), and increasingly decide who to kill and who to let live in combat warfare (Abraham, 2024; Asaro, 2012). It is crucial to emphasise that AI systems are not just a technology one uses, but are themselves a governance tool in public policy to establish, manage and enforce social orders. This pervasive form of government by algorithm, which Danaher (2016) coins 'algocracy', or Rouvroy and Berns (2013) refer to as 'algorithmic governmentality', shows a trend towards AI supporting or even replacing police, military, legislative and administrative action. Another trend in the embedding of AI is the dominance of social media platforms and marketplaces. There is a growing centralisation around commercial platforms that act as powerful providers, gatekeepers and bottlenecks for AI applications and services. Commercial platforms use AI technology to evaluate, sort and recommend information flows and users. In doing so, platforms pervasively reshape communication relationships and behaviour (Gillespie, 2010; Nitzberg and Zysman, 2022; Srnicek 2017). Through this central position, platforms reconfigure human-AI situatedness (Suchman, 2007), enforcing new modes of interaction, values, spatial and temporal experiences (e.g., intimacy, ubiquity, acceleration). In terms of trust, the use of AI in society, governance and platforms represents an important embedding that needs to be accounted for conceptually. With AI taking on key tasks in the operation and management of platforms, platforms themselves are also theorised as trust mediators (Bodó, 2021b). These virtual meeting places become sites of trust production by matching buyers and sellers, potential sex partners or bridging transactional uncertainties between customers. Undoubtedly, trust can be built here by platforms moderated by AI - but in turn, as Bodo (ibid.) puts it, it is crucial "to inquire whether we can trust technology to produce trust" (2680).

As shown in this section, trust in AI extends from the obvious and apparent AI application to a network of actors and ties. Moreover, it must also be understood as a governance tool for managing social orders, playing a central role in public policy and in the platformisation of widely used digital services. But: How can users control whether this network of relationships embedding AI is trustworthy? They cannot see or understand all the consequences of the specific technical and political choices made by all actors in the design of AI systems. Nor do they have the skills, let alone the information, to grasp whether AI systems are functioning properly and are integer (for example, by not producing biased results or spreading misinformation). In essence, policymakers must consider that users are being presented with an AI end product that remains completely closed and opaque in its design process, its operating mechanisms, and its underlying normative choices.

It seems intuitive that the much-hailed ethical principles of transparency and autonomy are an essential pillar of a TAI standard, at least to counter this myriad of complexity and opacity. However, much recent empirical research shows that evidence is complicated and not as intuitive as ethical guidelines might suggest (Felzmann et al., 2019). In a German study, König et al. (2022) show that in interaction with personal AI assistants users "do value explainable AI, i.e., high transparency of the AI assistants, [while] this feature barely offsets even a monthly price of 1.99 Euros as compared to no costs" (8). Moreover, Waldman and Martin (2022) show that AI transparency alone does not suffice to judge public policy decisions based on algorithms as legitimate, "countering arguments for greater transparency as a governance solution" (12). They suggest that a human in the decision-making loop is crucial for sensitive areas like policing or judiciary where it is perceived that human capacities and skills crucially matter, which is also supported by Lee (2018). But then again to the contrary, Krügel et al. (2022) show that human oversight does not counter user overconfidence in corrupted algorithms, transforming humans in the loop without digital literacy into "zombies in the loop" (1). While scholarship needs to further explore which arrangements of transparency and human oversight matter in AI contexts, it is already clear that it is not enough to disclose all the different actors and factors that make up the web of trust around an AI application. Realistically, policy makers need to consider that users cannot monitor this myriad and assess the trustworthiness of all actors. To provide TAI, it is essential that users can rely on institutional governance frameworks that establish, maintain and guarantee a trustworthy web of actors. Regulators and their governance role are central to bridging uncertainties. It is within their mandate and competence to implement a regulatory framework that creates systems of trust assurance.

Trust the AI regulatory framework. Governance ensnared between AI interest mediation and value trade-offs. The sociological and institutional literature on trust recognises for long that trust relationships rely on higher-order arrangements that bridge contexts of social uncertainty and knowledge gaps (Misztal, 1996; North, 1990; Sztompka, 1998; Zucker 1985). The complexity of managing different actors influencing TAI demonstrates both the importance and the challenge for public administrations dealing with AI. To date, AI governance modalities make use of both principle-based top-down regulation and market-based selfregulation, using a variety of cooperation and competition logics to govern AI. While the global AI governance landscape is still scattered and evolving, recently, the formation of more coercive regulatory regimes, most notably at the EU level with the AIA, DSA and Digital Market Act (DMA)[5] (European Commission, 2022a) come into being.

Before delving into policy details, it is important to take a step back and adopt the perspective of public administrations trying to establish trustworthiness for their AI regulatory frameworks and bridge the uncertainty faced by AI users. Their challenge is to manage and balance the different imperatives present in society. These include industry interests for a deregulatory capitalist agenda, the administrations' own internal security and geopolitical interests, while addressing users' concerns about AI and its alignment with existing legal norms and constitutional frameworks. All these imperatives follow different logics and engage with different narratives in the process of AI regulation, making it difficult to co-construct a common understanding of AI, let alone a consensus for appropriate policymaking (König et al., 2021). Recent special issues on the governance of AI (see Büthe et al., 2022; Taeihagh, 2021) have attempted to structure a still young field and aim to find a common language. Here I follow Büthe et al. (2022) that "laws, regulations and other measures to govern AI (…) do not so much reflect inherent characteristics or objective truths about the technology, but reflect political actors' perceptions given those actors' predisposition" (1722).

Instead of talking about different actors in the policy process, however, it is more appropriate to conceptualise the AI policy process as a bargaining field of conflicting players trying to maximise their stakes. This shift in perspective helps to understand the phenomenon of trust and distrust in AI arising from governance frameworks. It is manifested in decisions about value trade-offs that seem inevitable in AI regulation. Politics is caught in a mediating tension, as it has to accommodate different narratives and imperatives of interests that contradict each other in the policy process. The motif of an ensnared state facing a regulatory dilemma has long been propagated by conflict state theorists such as Offe (1972) or Alford and Friedland (1985), and is also present in the hegemony theory of Laclau (1996) and Mouffe (2013). Recent scholarship has aimed to reintroduce agonistic paradigms into technopolitics, mostly in opposition to a perceived dominant deliberative reading of politics in technology assessment (see discussion by Delvenne and Parotte, 2019; Schröder 2019). From an agonistic political perspective, administrations are pressured to consider different narratives and political interests - without taking sides - in order to be perceived as integer, legitimate and trustworthy. Favouring one societal imperative concerning AI (allowing ubiquitous access to user data to support the rise of AI start-ups) may neglect the concerns of another player (users' concerns about privacy and data autonomy) and undermine the trustworthiness of the administration. In this context, Sztompka (1998) paradoxically speaks of the need for an "institutionalized distrust" (1). After all, it is not surprising that conflicting opinions and interests clash around AI. On the positive side, it can also be read as a constitutive and vital element of democratic political culture. As Bodó (2021a) writes:

> "This competition for the autonomous powers of the state (…) requires the development of complex networks of institutional distrust, which reflect both the distrust among different societal groups with radically divergent and competing interests, and the very real possibility that any of these groups may overtake any of the bodies of the state" (12).

"Overtaking" may have a strong connotation, but issues of regulatory capture, clientelism and outright corruption pose a serious threat to public perceptions of AI regulation and political mandate. This threat is illustrated by the fact that AI regulation faces pervasive value trade-offs. If some stakeholders value a regulatory framework that promotes transparency, corporate accountability, user autonomy & privacy, and progressive fairness standards for vulnerable groups in AI applications, this comes with the caveat of reducing the efficiency and accuracy of those AI applications. For example, designing AI applications to be more explainable (higher interpretability) is time and cost-consuming. It also reduces the complexity of AI systems and curtails their performance output (Baryannis et al., 2019). Or, it has been shown that to make an AI less discriminatory, a programmer must suppress all correlations and proxies associated with a protected category, such as 'gender' or 'age'. This has a significant impact on making an AI model broader and less specific, further being complicated by different fairness principles inherently excluding each other (Kleinberg et al., 2016; Wong, 2020). Higher accuracy means better performance (algorithmic efficiency), but can also lead to disparate impact (more discrimination against vulnerable groups) (Barocas and Selbst, 2016). It goes without saying that higher standards of privacy and corporate accountability would be highly valued by many users, but would be at odds with large data-driven business models of big commercial platforms. Such inevitable trade-offs in AI governance represent an apple of discord, struggling for harder and softer AI regulation, with the risk of producing inconsistent or partisan regulatory frameworks. The EU's AI regulatory process is a case in point.

Recent reports by the 'Corporate Europe Observatory', 'Transparency International' and 'Euroactive' show how big Tech, corporate think tanks, and trade and business associations are active in blocking and watering down AI regulation in Brussels. Big Tech, largely dominated by US firms, have "spen[t] over € 97 million annually lobbying the EU institutions (…) ahead of pharma, fossil fuels, finance, or chemicals" (Bank et al., 2021: 6). In 2023 industry lobbyists had by far the most meetings with the EU commission on the AIA, featuring 86% of all behind closed-door meetings (73 out of 98 meetings), and were most active in agenda and standard setting (Corporate Europe Observatory, 2023; Kergueno et al., 2021). For the AIA "tech companies have reduced safety obligations, sidelined human rights and anti-discrimination concerns" (Schyns, 2023: 3). Leaked documents strikingly show how companies try to pressure policy makers for a deregulatory agenda by staging narratives like "Big tech is 'irreplaceable' when it comes to problem solving", "we're just defending SMEs and consumers", "Europe wins the tech race against China, or it falls back into the Stone Age" (Bank et al., 2021: 27). In the final round of discussions on the AIA, these lobbying efforts have been directed against the designation of general-purpose AI as a 'high risk' category in the AIA, with industry fearing that it would overburden and stifle innovation with strict conformity assessments. European startups like 'Mistral' and 'Aleph Alpha' teamed up with US big Tech companies and derailed, with direct ties to political executives in France or Germany, the policy-making process on the last meters. Industry managed to water down the binding fundamental right assessment proposed by the European Parliament on general-purpose AI into mere transparency rules (Corporate Europe Observatory, 2023; Hartmann, 2023).

Reports that show such a disproportionate favouring of industry interests can be a blow to public perceptions of AI. If users feel (and truly, a feeling may suffice) that regulation is being framed in such a way that AI regulatory trade-offs favour powerful interests but lack democratic integrity, they may be reluctant to trust it. Problematically, distrust can become diffuse and endemic - and then persistently damaging - when the contacts between policy and an interest group become too close and increasingly indistinguishable. Lobbying and partisan agenda-setting takes place behind the scenes. Unable to identify and address those responsible, some publics quickly direct their sentiments of distrust towards diffuse upper hierarchies such as 'the system', 'the powerful elites', 'those Eurocrats in Brussels'. The revolving door phenomenon can certainly fuel this perception. This is undoubtedly the case with AI at the EU regulatory level, as "three quarters of all Google and Meta's EU lobbyists have formerly worked for a governmental body at the EU or member state level" (Schyns, 2023: 7). In general, interest trade-offs are not necessarily problematic if regulator communication is transparent and honest. How value trade-offs are communicated and accommodated is an essential feature of managing expectations, hopes and fears around AI. It draws central attention to the discursive dimension of AI, which leads to the final analytical dimension that pairs trust with AI.

Promoting trustworthy AI through narratives: mediating meaning & attention. Trust in AI is strongly mediated by its discursive framing, which creates meaning what to expect from AI, the promises and fears it embodies, and the problems it is supposed to solve. Hence, the societal role which AI shall fulfil is not innate in technical details but is socially constructed and harnessed. Science and technology needs the social narrative to justify itself as valid, legitimate, needed, and strived for. As will be argued, TAI narratives have a dual societal function: they create acceptance, topicality and attract investments around AI, while at the same time silencing and bridging value conflicts and contradictions as assessed in the previous section.

AI is a technology that is very rich from a narrative standpoint. The extensive discursive embedding of AI with human concepts such as 'thinking', 'autonomy' and 'intelligence' shapes perceptions of AI in both public and expert domains. Since its beginning, AI has raised expectations and dreams of exuberant achievements, constantly entertaining the thought of outperforming the human (Campolo and Crawford, 2020; Dandurand et al., 2022; Natale and Ballatore, 2020). These narratives are often embedded in the binary of hopes and fears, or redemption or doom, most concretely embodied in fictional narratives around AI (Cave and Dihal, 2019). But the fictional quickly conflates into the real, with AI myths being echoed in public arenas shaping overall AI sense making (Crépel et al., 2021). Framed perceptions of AI raise expectations that may be frustrated if promises are not kept, negatively influencing perceptions of both the trust-giver (the communicator of promises, such as providers or regulators) and the then demystified AI systems. The often-exaggerated image that conveys the potential and danger of AI is critical for the realm of trust, as trust relationships are built on emotional expectations. When users are confronted with a discrepancy between exaggerated promises and the actual reality, this can lead to feelings of dishonesty, disappointment and even betrayal.

Given this context, empirical work shows how nation-states and supranational institutions have actively positioned themselves in the AI arena. Administrations portray themselves in an 'AI race' (Cave and ÓhÉigeartaigh, 2018), employing deterministic rhetoric of an 'inevitable' societal path towards AI. This future trajectory is fuelled by rhetorics of TINA (there is no alternative), politically surrendering to the logics of international economic competition. Likewise, societies being constantly shaken by the exhausting reality of crises transforms AI's role from a technology into that of a saviour, nourishing the epic tale of redeeming society from its current structural problems, such as the urban mobility crisis, social inequalities, or climate change. This solutionist aura (Morozov, 2013) that surrounds AI in the political and cultural realm reifies it as given and needed – thereby defining the toolkit to combat socially deeply rooted problems. With the race to AI portrayed as inevitable, a race to AI regulation (Smuha, 2021) is also evoked, pressuring governments to come up with effective regulatory frameworks. However, selling smart AI-based solutions while ignoring deep-rooted social problems can be a pitfall for TAI. The sociology of expectations and STS warn about the risk of such tech-ubiquity leading to path dependencies and lock-ins (Borup et al., 2006; van Lente and Rip, 1998). Managed public expectations of AI can easily turn into demands on governments. As I have argued elsewhere with Bareis and Katzenbach (2022), deconstructing the consistency of national AI imaginaries: "Once governments proclaim bold promises, they are on the spot to deliver and perform their capabilities" (874). The praise of technology talk becomes performative and can increase the pressure not to disappoint users. Stakeholders are playing with the trust of AI-users if raised expectations are not met and promises prove empty – or scandals shatter the before hailed AI solutions.

In general, not only AI but also TAI has become a buzzword in politics. As outlined in the section before, the EU has framed its entire regulatory framework with the emblem of TAI. While TAI remains completely under-defined, it functions as an empty signifier that has its political function. By deploying the TAI frame, the EU Commission can rhetorically accommodate stakeholders and their conflicting interests and unify a contested field of actors in a seemingly harmonised and consensual regulatory framework. From the outset, "AI industry can read 'trustworthiness' as a call for robustness, while ethicists and legal experts can simultaneously imagine that the document puts forward the agenda of making AI development more ethical and lawful" (Stamboliev and Christiaens, 2024:6). Thus, TAI functions as a unifier to bridge different interests, but this comes with a significant caveat: the carving out of what TAI actually entails. This semantic emptiness may even be cherished and promoted by political actors, but of course it would then lack any substance and meaningful content. Worse, if these empty signifiers are revealed as a strategy to obscure power structures in regulatory processes, the blow to TAI and AI governance bodies can be substantial.

Concluding remarks

Trustworthy AI (TAI) has recently been widely employed in the context of AI regulation and in ethical debates around AI. This paper aims to structure and advance the debate, doing justice to a complex socio-political phenomenon that has suffered from being reduced to a semantically carved-out buzzword. This paper argues that the actual requirements for linking trust and AI are demanding – but also rewarding. Rather than following the dominant path in AI research of linking trust to ethical principles such as fairness, transparency, or privacy, or to technical properties such as robustness, efficiency, or accuracy, I hope to have shown that the phenomenon of TAI (while certainly being influenced by these) mobilises larger epistemic and social dimensions. Any technical approach to de-biasing, auditing, or making AI more transparent has its merits, but ultimately falls short of capturing and doing justice to the variously situated realms that constitute TAI. These include a) AI as an intersubjective relationship, with trust being negotiated through AI as a quasi-other; b) the embedding of AI in a network of actors from programmers to platform gatekeepers; c) the regulatory role of governance in bridging trust uncertainties and deciding on AI value trade-offs; and d) the role of narratives and rhetoric in mediating AI and conflictual AI governance processes (see overview Table 1). Admittedly, the analytical scheme is a heuristic and therefore necessarily abstract. I have executed each dimension with regard to AI in this paper, but in reality, they easily conflate. Some work more in the foreground with interfaces and materialities, others are enmeshed and implicit in power-relations and hierarchies, or framed by conversations about AI Hollywood blockbusters or technical policy results. However, for policy makers and researchers, the analytical scheme has its merit as it structures a scattered debate, points to regulatory requirements and brings clarity for further research trajectories.

Given the regulatory perspective, first, one must state that there are clear policy gaps in the European regulatory acts (other international proposals are still in the making) like the AIA, DSA and DMS. This concerns a questionable self-assessment and third-party risk assessment approach, or insufficient accountability duties for the identification and labelling of AI-generated synthetic content on platforms and search engines. With synthetically generated content flooding the internet, there is an increasing societal disorientation to what extent the blurring of the authentic and factual with the fake and false is socially and politically acceptable. This especially concerns AI applications in fields where users are most vulnerable such as care, education or sexuality.

Second, recent scholarship around internet regulation theorized governance as an open-end reflexive coordination in a complex network of social ties, "ordering processes from the bottom-up rather than proceeding from regulatory structures" (Hofmann et al., 2017: 1413). This actor-network inspired governance perspective serves well to bring all actors who are involved in AI production and distribution to the foreground, but understates the very nature of power and political bargaining between these actors. Hofmann et al. state that governance, here understood as coordination, "becomes reflexive when ordinary interactions break down or become problematic" (ibid.: 1414). This implied deliberative take of governing a complex network would misconceive the nature of hierarchical politics, though. Rather than leaning on a reflexive notion of politcs, I have put forward an agonistic picture of

AI governance, depicting strivings for hegemony and agenda setting between players in deciding upon value trade-offs. This perspective serves to understand the political dimension in installing trust or provoking distrust in AI, tackling issues of regulatory capture or revolving doors. These phenomena, of course, are not only limited to AI but also emerge alongside other regulations. Not surprisingly, though, it is especially prevalent when big Tech is aiming to make big money.

Third, I indicated that the carving out of TAI may not only be the consequence of a scattered debate but also depicts political strategy. I have highlighted the role of discourses and narratives for trust in AI, managing expectations through playing around with hopes and fears. It is revealing that transparency, integrity and honesty have such a low standing in political processes. The fact that the implementation of AI involves value trade-offs is not the fault of policymakers - but the euphemised way in which it is presented, not to mention the unbalanced and hidden lobbying that is allowed to take place, certainly is. Every trade-off with AI has its benefits and perils for society, and these can and should be fully and transparently articulated – and publicly discussed. This would actually relieve politicians of much of the pressure to sugar-coat bad deals and spare them from manoeuvring themselves into rhetorical traps they then struggle to escape. Clarifying the stakes, the actors and their interests is in itself a transparency value that could substantially (re) build trust in political processes and, consequently, in their regulatory objects – in this case, AI.

By disentangling the relationship between trust and AI, this scholarship situates itself within the agenda of critical policy studies (Paul, 2022) and critical algorithm studies (Seaver, 2017). To successfully (dis-)integrate AI for the benefit of all, an understanding of how algorithmic phenomena shape, maintain and challenge society and its order is a pivotal precondition. This understanding calls for disciplinary transgression where needed to disclose how the technical is inscribed, mediated and practised in the social.

Acknowledgements

I would like to thank Christian Katzenbach and his colloquium, Julia Valeska Schröder and Clemens Ackerl for their helpful comments on earlier drafts of this paper. I am also grateful for three anonymous reviewers and Rocco Bellanova who did not lose trust in this interdisciplinary endeavor in various rounds of review.

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Jascha Bareis https://orcid.org/0000-0002-5823-0437

Notes

1. I follow Duenas-Cid and Calzati (2023) who argue that distrust is not the binary counterpart of trust, implying an opposite end of the same continuum. As they argue with regard to data-driven technologies, trust and distrust must be "regarded as independent yet complementary facets" that coexist (6) and "contribute together to their coming into being in different contexts" (14). Given the limited scope of the paper, I will mainly focus on the relationship of trust and AI but I will still prove their point and show how trust and distrust shape each other's realms and dynamics.
2. See the huge inventory of ethics guidelines by AlgorithmWatch https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/.
3. See systematic reviews and frameworks on the multiple ethical guidelines: Floridi & Cowls, 2021; Jobin et al., 2019.
4. Other countries like China or the US have also forwarded AI regulatory initiatives, like the 2023 Chinese "Interim Measures of the Management of Generative Artificial Intelligence Services" or the 2023 US executive order on "Safe, Secure and Trustworthy AI". Especially in the US executive order trustworthiness is stressed but also stays ill defined. I my analysis I will especially focus on the European regulatory AI framework as to date, it is the one which is most elaborated.
5. See footnote 4.
6. A first version of this scheme was developed together with Clemens Ackerl at the research group social trust in learning systems at ITAS, Karlsruhe.

References

Abraham Y (2024) 'Lavender': The AI machine directing Israel's bombing spree in Gaza. *+972 Magazine*. Accessed from: https://www.972mag.com/lavender-ai-israeli-army-gaza/.

Alford RR and Friedland R (1985) *Powers of Theory: Capitalism, the State, and Democracy*. Cambridge: Cambridge University Press.

Angwin J, Larson J, Mattu S, et al. (2016) Machine bias. Ethics of Data and Analytics, 254–264. Auerbach Publications.

Asaro P (2012) On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross* 94(886): 687–709.

Bader V and Kaiser S (2019) Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence. *Organization* 26(5): 655–672.

Baier A (1986) Trust and antitrust. *Ethics* 96(2): 231–260.

Bank M, Duffy F, Leyendecker V, et al. (2021) *The Lobby Network: Big Tech's Web of Influence in the EU*. Brussels: Corporate Europe Observatory. Retrieved from: https://corporateeurope.org/en/2021/08/lobby-network-big-techs-web-influence-eu.

Bareis J and Katzenbach C (2022) Talking AI into being: The narratives and imaginaries of national AI strategies and their performative politics. *Science, Technology, & Human Values* 47(5): 855–881.

Barocas S and Selbst AD (2016) Big data's disparate impact. *California Law Review* 104: 671.

Baryannis G, Dani S and Antoniou G (2019) Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. *Future Generation Computer Systems* 101: 993–1004.

Bodó B (2021a) The commodification of trust. Blockchain & Society Policy Research Lab Research Nodes, 1. DOI: 10.2139/ssrn.3843707.

Bodó B (2021b) Mediated trust: A theoretical framework to address the trustworthiness of technological trust mediators. *New Media & Society* 23(9): 2668–2690.

Borup M, Brown N, Konrad K, et al. (2006) The sociology of expectations in science and technology. *Technology Analysis & Strategic Management* 18(3–4): 285–298.

Botsman R (2017) *Who Can You Trust? How Technology Brought Us Together–And Why It Could Drive Us Apart*. London: Penguin UK.

Büthe T, Djeffal C, Lütge C, et al. (2022) Governing AI – attempting to herd cats? Introduction to the special issue on the governance of artificial intelligence. *Journal of European Public Policy* 29(11): 1721–1752.

Campolo A and Crawford K (2020) Enchanted determinism: Power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society* 6(0): 1–19.

Cave S and Dihal K (2019) Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* 1(2): 74.

Cave S and ÓhÉigeartaigh SS (2018) An AI race for strategic advantage: rhetoric and risks. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 36-40).

Chesney B and Citron D (2019) Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review* 107: 1753.

Coeckelbergh M (2012) Can we trust robots? *Ethics and Information Technology* 14(1): 53–60.

Coleman JS (1986) Social theory, social research, and a theory of action. *American Journal of Sociology* 91(6): 1309–1335.

Corporate Europe Observatory (2023) Byte by byte. How big Tech undermined the AI Act. Accessed from: https://corporateeurope.org/en/2023/11/byte-byte.

Crépel M, Do S, Cointet JP, et al. (2021) Mapping AI Issues in Media Through NLP Methods. In *CHR* (pp. 77-91).

Danaher J (2016) The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology* 29: 245–268.

Dandurand G, Blottière M, Jorandon G, et al. (2022) Training the News: Coverage of Canada's AI Hype Cycle (2012–2021). INRS - Urbanisation Culture Société.

Delvenne P and Parotte C (2019) Breaking the myth of neutrality: Technology assessment has politics, technology assessment as politics. *Technological Forecasting and Social Change* 139: 64–72.

Duenas-Cid D and Calzati S (2023) Dis/Trust and data-driven technologies. *Internet Policy Review* 12(4): 1–23.

EU High-Level Expert Group on AI (2019) Ethics Guidelines for Trustworthy Artificial Intelligence. Accessed from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

European Commission (2020a) Assessment List for Trustworthy AI (ALTAI). Accessed from: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

European Commission (2020b) On Artificial Intelligence – A European approach to excellence and trust. Accessed from: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

European Commission (2022a) Digital Markets Act (DMA). Accessed from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1925.

European Commission (2022b) Digital Service Act (DSA). Accessed from: https://eur-lex.europa.eu/legal-content/EN/TXT/?toc=OJ%3AL%3A2022%3A277%3ATOC&uri=uriserv%3AOJ.L_.2022.277.01.0001.01.ENG.

European Commission (2024) Artificial Intelligence Act (AIA). Accessed from: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A52021PC0206.

Felzmann H, Villaronga EF, Lutz C, et al. (2019) Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society* 6(1). DOI: 2053951719860542.

Floridi L and Cowls J (2021) A unified framework of five principles for AI in society. *Harvard Data Science Review* 1(1). DOI: 10.1162/99608f92.8cd550d1.

Folberth A, Jahnel J, Bareis J, et al. (2022) Tackling problems, harvesting benefits–A systematic review of the regulatory debate around AI. arXiv preprint arXiv:2209.05468.

Gillespie T (2010) The politics of 'platforms'. *New Media & Society* 12(3): 347–364.

Gorwa R, Binns R and Katzenbach C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7(1).

Güzeldere G and Franchi S (1995) Dialogues with colorful "personalities" of early AI. *Stanford Humanities Review* 4(2): 161–169.

Hardin R (2002) *Trust and Trustworthiness*. New York: Russell Sage Foundation.

Hartmann T (2023) AI Act: French government accused of being influenced by lobbyist with conflict interests. www.euractiv.com. https://www.euractiv.com/section/artificial-intelligence/news/ai-act-french-government-accused-of-being-influenced-by-lobbyist-with-conflict-of-interests/.

Hoffmann R (2019) Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22(7): 900–915.

Hofmann J, Katzenbach C and Gollatz K (2017) Between coordination and regulation: Finding the governance in internet governance. *New Media & Society* 19(9): 1406–1423.

Ihde D (2009) *Postphenomenology and Technoscience. The Peking University Lectures*. Peking: Peking University Press.

Jobin A, Ienca M and Vayena E (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9): 389–399.

Jones K (1996) Trust as an affective attitude. *Ethics* 107: 4–25.

Kaur D, Uslu S, Rittichier KJ, et al. (2023) Trustworthy artificial intelligence: A review. *ACM Computing Surveys* 55(2): 1–38.

Kergueno R, Aiossa N, Pearson L, et al. (2021) Deep pockets, open doors: Big tech lobbying in Brussels. *Transparency International EU*: 1–21.

Kleinberg J, Mullainathan S and Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.

König H, Baumann MF and Coenen C (2021) Emerging technologies and innovation—hopes for and obstacles to inclusive societal co-construction. *Sustainability* 13(23): 13197.

König PD, Wurster S and Siewert MB (2022) Consumers are willing to pay a price for explainable, but not for green AI. Evidence from a choice-based conjoint analysis. *Big Data & Society* 9(1): 205395172110696.

Krafft PM, Young M, Katell M, et al. (2020) Defining AI in policy versus practice. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 72-78.

Krarup T and Horst M (2023) European artificial intelligence policy as digital single market making. *Big Data & Society* 10(1): 20539517231153811.

Krügel S, Ostermaier A and Uhl M (2022) Zombies in the loop? Humans trust untrustworthy AI-advisors for ethical decisions. *Philosophy & Technology* 35(1): 611.

Laclau E (1996) *Emancipation(s)*. New York: Verso.

Latour B (1994) On technical mediation. *Common Knowledge* 3(2): 29–64.

Laurim V, Arpaci S, Prommegger B, et al. (2021) Computer, whom should I hire? Acceptance criteria for artificial intelligence in the recruitment process. Proceedings of the 54th Hawaii international conference on system sciences, p. 5495.

Lee MK (2018) Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5(1). DOI: 10.1177/2053951718756684.

Loh J (2019) *Maschinenethik und Roboterethik. Handbuch Maschinenethik*. Heidelberg: Springer, 75–93.

Luhmann N (1988) Familiarity, confidence, trust: Problems and alternatives. In: Gambetta D (eds) *Trust: Making and Breaking Cooperative Relations*. Oxford: Basil Blackwell, 94–107.

Mackenzie A (2015) The production of prediction: What does machine learning want? *European Journal of Cultural Studies* 18(4-5): 429–445.

McKnight DH, Choudhury V and Kacmar C (2002) Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research* 13(3): 334–359.

McLeod C (2002) *Self-Trust and Reproductive Autonomy*. Cambridge: MIT Press.

McLeod C (2021) Trust. The Stanford Encyclopedia of Philosophy (Fall 2021 Edition). Available at: https://plato.stanford.edu/archives/fall2021/entries/trust/.

McMillan L and Varga L (2022) A review of the use of artificial intelligence methods in infrastructure systems. *Engineering Applications of Artificial Intelligence* 116: 105472.

Misztal BA (1996) *Trust in Modern Societies: The Search for the Bases of Social Order*. Cambridge: Polity Press.

Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1(11): 501–507.

Morozov E (2013) *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: Public Affairs.

Mouffe C (2013) *Agonistics: Thinking the World Politically*. London: Verso.

Munn L (2022) The uselessness of AI ethics. AI and Ethics, 1-9.

Natale S and Ballatore A (2020) Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence: The International Journal of Research into New Media Technologies* 26(1): 3–18.

Nickel PJ (2007) Trust and obligation-ascription. *Ethical Theory and Moral Practice* 10(3): 309–319.

Nickel PJ (2013) Trust in technological systems. In: de Vries MJ, Hansson SO and Meijers AWM (eds) *Norms in Technology*. Dordrecht: Springer, 223–237.

Nickel PJ, Franssen M and Kroes P (2010) Can we make sense of the notion of trustworthy technology? *Knowledge, Technology & Policy* 23(3): 429–444.

Nitzberg M and Zysman J (2022) Algorithms, data, and platforms: The diverse challenges of governing AI. *Journal of European Public Policy* 29(11): 1753–1778.

North DC (1990) *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.

OECD (2019) OECD AI Principles - Recommendation of the Council on Artificial Intelligence. Accessed from: https://oecd.ai/en/ai-principles.

Offe C (1972) *Strukturprobleme des kapitalistischen Staates*. Frankfurt a. M.: Suhrkamp.

Páez A (2019) The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines* 29(3): 441–459.

Paul R (2022) Can critical policy studies outsmart AI? Research agenda on artificial intelligence technologies and public policy. *Critical Policy Studies* 16(4): 497–509.

Paul R (2023) European artificial intelligence "trusted throughout the world": Risk-based regulation and the fashioning of a competitive common AI market. *Regulation & Governance*. DOI: 10.1111/rego.12563.

Radu R (2021) Steering the governance of artificial intelligence: National strategies in perspective. *Policy and Society* 40(2): 178–193.

Reinhardt K (2022) Trust and trustworthiness in AI ethics. AI and Ethics, 1-10.

Rouvroy A and Berns T (2013) Gouvernementalité algorithmique et perspectives d'émancipation: Le disparate comme condition d'individuation par la relation? *Réseaux* 177: 163–196.

Russell S and Norvig P (2022) Artificial Intelligence: A Modern Approach, 4th Global ed.

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

Salles A, Evers K and Farisco M (2020) Anthropomorphism in AI. *AJOB neuroscience* 11(2): 88–95.

Scheutz M and Arnold T (2016, March) Are we ready for sex robots? In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 351-358). IEEE.

Schröder JV (2019) Das Politische in der Technikfolgenabschätzung: Reflexionen mit der pluralen, radikalen Demokratietheorie von Laclau und Mouffe. *TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis/Journal for Technology Assessment in Theory and Practice* 28(3): 62–67.

Schyns C (2023) The lobbying ghost in the machine. Corporate Europe Observatory. Brussels, Belgium. Accessed from: https://corporateeurope.org/en/2023/02/lobbying-ghost-machine.

Seaver N (2017) Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4(2). DOI: 10.1177/2053951717738104.

Sheridan TB (2020) A review of recent research in social robotics. *Current Opinion in Psychology* 36: 7–12.

Silva de Barcelos A, Gomes MM, da Costa CA, et al. (2020) Intelligent personal assistants: A systematic literature review. *Expert Systems with Applications* 147: 113193.

Simion M and Kelp C (2023) Trustworthy artificial intelligence. *Asian Journal of Philosophy* 2(1): 8.

Smuha NA (2021) From a 'race to AI' to a 'race to AI regulation': Regulatory competition for artificial intelligence. *Law, Innovation and Technology* 13(1): 57–84.

Söllner M, Hoffmann A and Leimeister JM (2016) Why different trust relationships matter for information systems users. *European Journal of Information Systems* 25(3): 274–287.

Srnicek N (2017) *Platform Capitalism*. London: Polity.

Stamboliev E and Christiaens T (2024) How empty is trustworthy AI? A discourse analysis of the ethics guidelines of trustworthy AI. *Critical Policy Studies* 1–18. DOI: 10.1080/19460171.2024.2315431.

Suchman L (2007) *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge: Cambridge University Press.

Suchman L (2023) The uncontroversial 'thingness' of AI. *Big Data & Society*, 10(2). DOI: 10.1177/20539517231206794.

Suchman L and Weber J (2016) Human-machine autonomies. In: Bhuta N, Beck S, Geis R, et al. (eds) *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge: Cambridge University Press, 75–102.

Sztompka P (1998) Trust, distrust and two paradoxes of democracy. *European Journal of Social Theory* 1(1): 19–32.

Taeihagh A (2021) Governance of artificial intelligence. *Policy and Society* 40(2): 137–157.

van Huijstee M, van Boheemen P and Das D (2021) Tackling deepfakes in European policy.

van Lente H and Rip A (1998) Expectations in technological developments: An example of prospective structures to be filled in by agency. In: Disco C and Meulen B (eds) *Getting New Technologies Together: Studies in Making Sociotechnical Order*. Berlin, Germany: Walter de Gruyter, 203–230.

Veale M, Matus K and Gorwa R (2023) AI and Global Governance: Modalities, Rationales, Tensions. *Annual Review of Law and Social Science* 19.

von Eschenbach WJ (2021) Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology* 34(4): 1607–1622.

Waldman A and Martin K (2022) Governing algorithmic decisions: The role of decision importance and governance on perceived legitimacy of algorithmic decisions. *Big Data & Society* 9(1). DOI: 10.1177/20539517221100449.

Wallach W and Allen C (2008) *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Warren ME (ed.) (1999) *Democracy and Trust*. Cambridge: Cambridge University Press.

Washington Post (2020, February 13) The suspicious video that helped spark an attempted coup in Gabon | The Fact Checker [Video]. YouTube. https://www.youtube.com/watch?v=F5vzKs4z1dc.

Wong P (2020) Democratizing algorithmic fairness. *Philosophy & Technology* 33(2): 225–244.

Zednik C (2019) Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology* 34: 265–288.

Zucker LG (1985) Production of trust: Institutional sources of economic structure, 1840 to 1920. In: Cummings LL and Staw B (eds) *Research in Organizational Behavior*. Greenwich, CT: JAI Press, 53–110.