AI & SOCIETY (2023) 38:733–745
https://doi.org/10.1007/s00146-022-01473-4

OPEN FORUM

Trust and ethics in AI

Hyesun Choung[1] · **Prabu David**[1] · **Arun Ross**[1]

Received: 31 July 2021 / Accepted: 13 April 2022 / Published online: 20 May 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract
With the growing influence of artificial intelligence (AI) in our lives, the ethical implications of AI have received attention from various communities. Building on previous work on trust in people and technology, we advance a multidimensional, multilevel conceptualization of trust in AI and examine the relationship between trust and ethics using the data from a survey of a national sample in the U.S. This paper offers two key dimensions of trust in AI—human-like trust and functionality trust—and presents a multilevel conceptualization of trust with dispositional, institutional, and experiential trust each significantly correlated with trust dimensions. Along with trust in AI, we examine perceptions of the importance of seven ethics requirements of AI offered by the European Commission's High-Level Expert Group. Then the association between ethics requirements and trust is evaluated through regression analysis. Findings suggest that the ethical requirement of societal and environmental well-being is positively associated with human-like trust in AI. Accountability and technical robustness are two other ethical requirements, which are significantly associated with functionality trust in AI. Further, trust in AI was observed to be higher than trust in other institutions. Drawing from our findings, we offer a multidimensional framework of trust that is inspired by ethical values to ensure the acceptance of AI as a trustworthy technology.

Keywords Trust in AI · Ethics requirements of AI · Public perceptions of AI

1 Introduction

The rapid growth of artificial intelligence (AI) technologies and their pervasive influence in our lives have heightened the importance of ethics, values, and a human-centric approach to the design and development of AI (Floridi et al. 2018). Modern AI is defined as "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy" (OECD 2019, pp. 23–24). Capabilities such as decision-making and autonomy differentiate AI from traditional technologies, such as cars or computer software, which work within protocols that are under human control. AI's autonomous functioning has changed the balance of power between humans and machines, requiring humans to trust technology. Moreover, deep learning algorithms that drive current AI lack transparency and explainability (Shin 2021), which adds to the challenge of creating trustworthy technology.

Trust is an essential construct in human relationships with an extensive body of literature (Rotenberg 2019). Human trust in technology (Hancock et al. 2011), particularly automation and AI, also has engendered significant interest in the academic community. Researchers believe trust is a fundamental step toward the social acceptance of new and potentially disruptive technologies (Wu et al. 2011). However, unlike other technologies, AI presents unique challenges for researchers because of its manifestations, such as anthropomorphic features, natural language processing, affective computing, and conversational abilities. These capabilities necessitate a new socio-technical understanding of trust that extends beyond mere functionality to human-like characteristics of the machine (Choung et al. 2022). In this emerging understanding, attributes such as concern for the well-being of society, benevolence, helpfulness, fairness and compassion have become necessary components of trust in human-AI interactions (Thiebes et al. 2021).

✉ Hyesun Choung
choungh@msu.edu

Prabu David
pdavid@msu.edu

Arun Ross
rossarun@cse.msu.edu

[1] Michigan State University, East Lansing, MI, USA

To develop trustworthy AI systems, various organizations and even governments have proposed ethics guidelines for AI (Jobin et al. 2019; Hagendorff 2020). These guidelines encourage fairness and promote AI applications that are unbiased, non-discriminatory, and beneficial to society. Special care is given to the unintended consequences of AI and its effects on vulnerable populations (Borgesius 2018). Although ethical values are actively promoted by AI technologists, it is not clear to what extent they influence trust among users. Therefore, in this study, we examine the importance assigned by the general population to ethical values of AI and the influence of these values on trust.

Specifically, we develop a multidimensional, multilevel understanding of trust using a survey of a representative sample of participants in the United States ($N = 525$). Through regression analysis, we explore the effects of demographics, dispositional trust, institutional trust, and familiarity with AI on the two key dimensions of trust. In the last step, we examine the perceived importance of the ethics requirements of AI among users and the effect of these requirements on trust perceptions.

2 Trust in AI: a multidimensional and multilevel approach

Trust is a fundamental human mechanism required to cope with vulnerability, uncertainty, complexity, and ambiguity in situations that collectively constitute a risk (Colquitt et al. 2007). As we begin to realize the potential of AI, to ensure continued success it is essential to understand the social and psychological mechanisms of trust in human-AI interaction (Gillath et al. 2021). Trust is defined as "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another" (Rousseau et al. 1998, p. 395). Traditionally, trust is tied to relationships between people and is required to build mutuality and interdependence between parties in human communication. In addition, trust is a multidimensional concept that is associated with trustees' characteristics, intentions, and behaviors (Lee and See 2004; Schoorman et al. 2007).

Mayer et al. (1995) define trust in humans as an amalgam of one's belief in another's ability, benevolence, and integrity. Ability refers to skills and competencies to successfully complete a given task. Benevolence pertains to whether the trustee has positive intentions that are not based purely on self-interest. And integrity describes the trustee's sense of morality and justice, such that the trusted party's behaviors are consistent, predictable, and honest. This three-dimensional understanding of trust is vital to interpersonal interactions, as it contributes to reliability and integrity in our dealings with fellow humans.

Past studies have extended the concept of interpersonal trust to human-technology relationships (Calhoun et al. 2019), especially to technology with human-like characteristics (Gillath et al. 2021). But some studies have noted that human users have different expectations with technologies and that the principles of interpersonal trust are not directly applicable to human-to-machine trust (e.g., Madhavan and Wiegmann 2007). McKnight et al. (2011) explained that trust in technology is qualitatively different from trust in people primarily because humans are moral agents, whereas technology lacks volition and moral responsibility. They offered three analogous dimensions for trust in technology—functionality, reliability, and helpfulness—to replace ability, integrity, and benevolence. Functionality refers to the capability of the technology, which the authors likened to human ability. Reliability is the consistency of operation, which is like integrity, and helpfulness indicates whether the specific technology is useful to users and is comparable to the benevolence dimension of human trust.

Although AI is a technology, it is involved in replacing or augmenting humans in tasks and decisions. In that sense, it is more than simply a technology, and McKnight's definition and the dimensions of trust in technology may require adjustment when applied to AI. Unlike traditional technologies that rely on user input and the execution of rules programmed by humans, AI has more autonomy. Further, AI's technical abilities and human-like characteristics create an interesting duality. For technology applications with greater humanness, a trust-in-humans scale was better at predicting relevant outcomes than a trust-in-technology scale (Lankton et al. 2015), which calls for a conceptualization of trust that combines the technology and in its human-like characteristics.

Though a conceptual definition of trust in AI is still evolving, it appears that at least two dimensions of AI, functionality and human-like characteristics, are relevant (Thiebes et al. 2021). Drawing from previous work, we focused on human-like trust in AI (benevolence and integrity) and functionality trust in AI (Mayer et al. 1995; Mcknight et al. 2011). The former dimension pertains to the social and cultural values of the algorithms and the values and ethics that undergird the design of AI technology. The latter dimension relates to the reliability, competency, expertise, and robustness of the technology. For human-like trust in AI, the trust is in the AI agent or system itself and not in the specific actions or operations (Choung et al. 2022).

2.1 Trust propensity and trust in institutions

Trust in the human-like attributes of AI is a holistic conceptualization that is subject to influences across levels of analysis, from the individual (i.e., intrapersonal and interpersonal) to the collective (i.e., institutions and society) (Fulmer

and Dirks 2018). This multilevel framework of trust furthers our understanding of the process through which trust is built and calibrated over time (Hoff and Bashir 2015). Therefore, this study incorporates dispositional trust (propensity to trust others) and institutional trust (trust in institutions).

Propensity to trust, or dispositional trust, is the general tendency to trust another person (Mayer et al. 1995). People hold different levels of propensity to trust, and it is a relatively stable disposition. Researchers have found that the propensity to trust another person predicts initial trustworthiness, which is the perception of how trustworthy another person is (Colquitt et al. 2007; Alarcon et al. 2018). Another trust-building process that may apply to AI is institution-based trust. Trust in institutions is at an all-time low, exacerbated by mis- and disinformation (Fulmer and Dirks 2018; Edelman 2021). In this study, we predict that trust propensity and trust in institutions are predictors of trust in AI.

H1: Trust propensity (dispositional trust) and trust in institutions (institutional trust) are positively associated with trust in AI.

2.2 Familiarity

While trust propensity and trust in institutions influence initial trust, trust changes over time based on familiarity and knowledge (Gefen et al. 2003). Familiarity reduces uncertainties and counteracts concerns based on reliance on past experience (Gulati 1995), and contributes to experiential trust, which is trust built on experiences. Familiarity builds trust by creating an appropriate context to predict and interpret the other party's behavior (Chen and Dhillon 2003). Furthermore, more familiarity implies greater accumulated knowledge derived from previous interactions, which leads to higher levels of trust (Gefen 2000). Based on these findings, the following hypothesis is derived.

H2: Familiarity (experiential trust) is positively related to trust in AI.

3 Ethics principles of AI

If trust in technology is partly dependent on human characteristics such as integrity and benevolence, then values and ethics take on added importance. Ethics in AI is a socio-technical challenge that demands the appropriate balance between the benefits of the technology and social norms and values (Chatila and Havens 2019). Although AI promises significant benefits to human life and well-being, experts from various perspectives concur that thoughtful consideration of fairness, accountability, and transparency of the technology is critical to developing trustworthy AI (Shin and Park 2019) and ensuring a sustainable society (Arogyaswamy 2020). Ethical AI enables the flourishing of all members of a society, including vulnerable populations, such as children, the elderly, and those with less power in the social hierarchy (Torresen 2018). For example, AI must be equally aware of the rights of employees while creating benefits for employers. Similarly, the rights of consumers, such as their privacy, must be considered alongside benefits to corporations.

The role of AI, then, is more than finding the appropriate solution within a specific domain. It includes evaluating outcomes and solutions offered by AI within societal values, which vary considerably by society. Furthermore, societal values are constantly in flux, requiring a framework that is flexible and can evolve over time. Recognizing these challenges, ethicists, thought leaders, nations and corporations have issued ethics guidelines (Jobin et al. 2019). Despite political and ideological differences between countries, as a human race, we share common values of the rights of individuals, justice, and common good, that serve as cross-cutting themes across proposals.

An analysis (Hagendorff 2020) of ethics guidelines from Google, Microsoft, and IBM, Organization for Economic Co-operation and Development (OECD) and Institute of Electrical and Electronics Engineers (IEEE), as well as the governments of China, the United States, and the European Union (EU), has identified common ethical requirements for AI. These include privacy, human agency, transparency, explainability, safety and cybersecurity, to name a few. A succinct framework of ethics requirements can be found in the guidance offered by the European Commission's High Level Expert Group on AI (AI HLEG) (2019) (See Table 1 for a summary).

The framework offered by the AI HLEG is particularly relevant for this study because it connects the trustworthiness of AI with ethics, which is the crux of the research reported in this paper. Drawing from fundamental human rights, such as respect for human autonomy, prevention of harm, fairness and explicability, the EU framework offers seven ethics requirements, which are summarized in Table 1: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination, and fairness; social and environmental well-being; and accountability. It is important to note that some of these ethical values and requirements may be at odds with one another, which the authors acknowledge. For example, improvement in safety in autonomous vehicles may require relinquishing more control to AI, which may be antithetical to the value of human autonomy. Such examples of competing goods and ethical tensions are common in AI applications in domains such as health, education, or law enforcement.

Despite the widespread understanding in the AI community of the importance of such moral dilemmas and ethics

Table 1 Seven ethics requirements for trustworthy AI proposed by AI HLEG

Ethics requirements for trustworthy AI	Description
1.Human agency and oversight	AI systems should allow people to make informed decisions. There should be a human oversight mechanism through a "human-in-the-loop" approach
2.Technical robustness and safety	AI systems should be safe, reliable, and reproducible to minimize unintended harm
3.Privacy and data governance	Ensure privacy and data protection, which requires an adequate data governance framework
4.Transparency	AI systems and business models should be transparent, and the AI systems' decisions should be explainable to the stakeholders. People need to be informed about the systems' capabilities and limitations
5.Diversity, non-discrimination, and fairness	AI systems should be accessible to all, and unfair biases should be avoided. Minimizing algorithmic bias is also important
6.Societal and environmental well-being	AI systems should benefit human beings and they should take into account the social impact and environmental consequences
7.Accountability	Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes

requirements, it is not clear how the public perceives them and what influence, if any, these requirements have on trust in AI. To examine the perceived importance of the ethical requirements of AI and their relationship to trust, the following two research questions are proposed. The influence of ethics requirements is examined using regression after controlling for demographics as well as dispositional, institutional, and experiential trust.

RQ1: What is the perceived importance among the public of the ethical requirements of AI?
RQ2: What influence, if any, do the ethical requirements of AI have on trust in AI?

4 Method

4.1 Participants and procedure

Data were collected from a general U.S. population using an online survey conducted in April 2021 through Qualtrics. A quota sampling method was used to ensure the representativeness of participants. A total of 525 respondents (age $M = 45.43$ $SD = 17.83$, 50.1% women, 65.9% White, 12% Black or African American, 12% Hispanic, 5.9% Asian, 4.4% other) participated in the study. The average time for survey completion was 10 min.

4.2 Measures

Survey items and the reliability of the scales are presented in the appendix. The survey was administered to current and potential consumers of AI products, and they were asked to evaluate smart technologies in general, such as smart home products (e.g., Google Home, Ring) and voice assistants (e.g., Siri, Alexa).

Dispositional trust or trust propensity was measured with three items (Frazier et al. 2013): "I usually trust people until they give me a reason not to trust them," "I generally give people the benefit of the doubt when I first meet them," and "My typical approach is to trust new acquaintances until they prove I should not trust them." Institutional trust was the mean of trust in three institutions—the federal government, corporations, and big technology companies. Experiential trust was calculated by averaging frequency of use of AI consumer products (smart home devices, smart speakers, virtual assistants, and wearable devices). In the last section of the survey, respondents rated the importance of each of the seven ethical requirements proposed by AI HLEG.

The items to measure trust in AI were constructed to represent the three pillars of the construct used in trust in humans (Mayer et al. 1995) and trust in technology (Mcknight et al. 2011): benevolence/helpfulness, integrity/reliability, and competence/functionality. Based on Choung et al.'s (2022) factor analysis, which yielded a two-factor structure, the items measuring benevolence and integrity were grouped together, and competence was treated as another dimension. The first dimension comprises human-like trust in AI (six items), and the second dimension comprises functionality trust in AI (five items).

4.3 Analytic approach

After examining the frequencies and means of public perceptions of AI, we used hierarchical regression for the analyses. The regression analyses examined human-like trust and functionality trust in AI as separate outcomes, and the predictor variables were entered in four stages. Respondents' age, gender, education, and income levels were entered in the first stage as control variables. Trust propensity and levels of trust in institutions were entered at stage two. Familiarity with AI technologies was entered at stage three. Importance

Table 2 Intercorrelations, means, standard deviations and ranges of variables used in regression analysis ($N = 525$)

Variable	2	3	4	5	6	7	8	9	10	11	12	Mean (SD)
1. Trust propensity	0.42	0.26	0.37	0.36	0.34	0.34	0.36	0.39	0.37	0.39	0.42	3.75 (0.97)
2. Trust in institutions	1	0.52	0.21	0.16	0.15	0.14	0.20	0.30	0.18	0.60	0.48	2.94 (1.20)
3. AI familiarity		1	0.17	0.14	0.17	0.19	0.19	0.25	0.16	0.55	0.46	2.70 (1.36)
4. Privacy and data governance[a]			1	0.68	0.69	0.64	0.52	0.50	0.52	0.27	0.34	3.93 (1.09)
5. Human agency and oversight[a]				1	0.70	0.68	0.56	0.52	0.56	0.27	0.34	3.84 (1.04)
6. Technical robustness and safety[a]					1	0.69	0.58	0.57	0.63	0.28	0.38	3.89 (1.04)
7. Transparency[a]						1	0.60	0.50	0.61	0.26	0.32	3.97 (1.02)
8. Diversity, non-discrimination and fairness[a]							1	0.65	0.57	0.24	0.35	3.91 (1.11)
9. Societal and environmental well-being[a]								1	0.59	0.36	0.40	3.77 (1.13)
10. Accountability[a]									1	0.27	0.37	4.03 (1.04)
11. Human-like trust in AI										1	0.75	3.27 (1.03)
12. Functionality trust in AI											1	3.64 (0.92)

All correlation coefficients are all significant at the $p = 0.01$ level. [a] One of seven ethical requirements presented by the European Commission's AI High Level Expert Group (AI HLEG 2019)

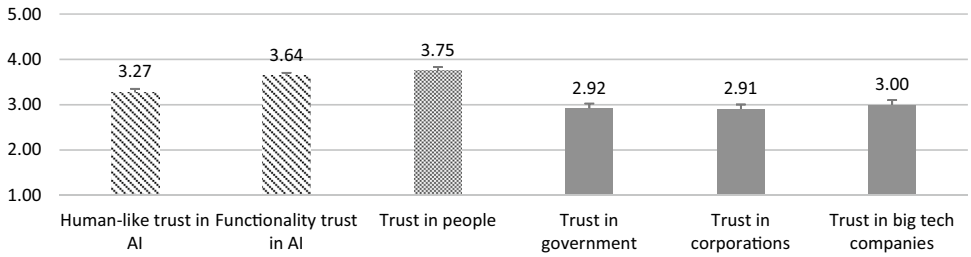

Fig. 1 Level of trust in AI

ratings of the seven ethical requirements for AI were entered together at stage four. Intercorrelations, means, and standard deviations for the independent and dependent variables are presented in Table 2. All correlations were statistically significantly, and the predictor variables were moderately correlated with the dependent variables.

5 Results

5.1 Levels of trust in AI

Overall, people held greater functionality trust in AI ($M = 3.64$, $SD = 0.92$) than trust in the human-like characteristics of AI ($M = 3.27$, $SD = 1.03$) (see Fig. 1). The levels of trust in AI were lower than general trust in other people ($M = 3.75$, $SD = 0.97$) but greater than trust in institutions ($M = 2.94$, $SD = 1.20$). This indicates that AI technologies are currently well-trusted and maintaining trust in AI requires shoring up trust in institutions, including Big Tech companies.

5.2 Predictors of trust in AI

Results from regression analyses (Tables 3, 4) showed that age and education were significant predictors of the level of human-like trust (age: $\beta = -0.22$, $p < 0.001$; education: $\beta = 0.25$, $p < 0.001$) and functionality trust in AI (age: $\beta = -0.17$, $p < 0.001$; education: $\beta = 0.20$, $p < 0.001$). Younger adults with higher education levels exhibited greater trust in AI. In addition, level of income was a significant positive predictor of the functionality trust in AI ($\beta = 0.12$, $p < 0.05$).

H1 postulated that propensity to trust and trust in institutions would positively predict trust in AI, after controlling for the demographics. As predicted, general trust propensity was a significant predictor of both human-like trust in AI ($\beta = 0.23$, $p < 0.001$) and functionality trust in Al ($\beta = 0.31$, $p < 0.001$). Trust in institutions also positively predicted human-like trust ($\beta = 0.44$, $p < 0.001$) and functionality trust in AI ($\beta = 0.29$, $p < 0.001$). Trust propensity and institutional trust together explained an additional 28% of the variance in human-like trust in AI (F (2, 518) = 127.34, $p < 0.001$) and 22% of the functionality trust in AI (F (2, 518) = 85.62, $p < 0.001$). Therefore, H1 was supported.

Introducing familiarity with AI technologies explained an additional 6% of the variance in human-like trust in AI

Table 3 Summary of hierarchical regression analyses for variables predicting human-like trust in AI ($N = 525$)

Variable	β	t	sr^2	R	R^2	ΔR^2
Step 1				0.37	0.14	0.14
Age	– 0.22	– 5.36***	0.048			
Gender (female = 1, male or other = 2)	0.07	1.51	0.004			
Education	0.25	4.88***	0.040			
Income	0.04	0.83	0.001			
Step 2				0.65	0.42	0.28
Age	– 0.17	– 4.64***	0.024			
Gender	0.03	0.78	0.001			
Education	0.13	3.06**	0.010			
Income	– 0.05	1.24	0.002			
Trust propensity	0.23	5.89***	0.039			
Trust in institutions	0.44	10.82***	0.131			
Step 3				0.70	0.48	0.06
Age	– 0.07	– 2.02*	0.004			
Gender	0.01	0.27	0.000			
Education	0.08	1.87	0.003			
Income	– 0.10	– 2.35*	0.006			
Trust propensity	0.19	5.05***	0.026			
Trust in institutions	0.33	8.18***	0.067			
AI familiarity	0.33	7.81***	0.061			
Step 4				0.71	0.51	0.02
Age	– 0.08	– 2.15*	0.004			
Gender	0.02	0.66	0.000			
Education	0.06	1.55	0.002			
Income	– 0.12	– 2.86**	0.008			
Trust propensity	0.14	3.45**	0.012			
Trust in institutions	0.33	8.05***	0.063			
AI familiarity	0.32	7.60***	0.056			
Privacy and data governance	– 0.02	– 0.40	0.000			
Human agency and oversight	0.09	1.69	0.003			
Technical robustness and safety	0.06	1.06	0.001			
Transparency	– 0.00	– 0.08	0.000			
Diversity, non-discrimination and fairness	– 0.08	– 1.64	0.003			
Societal and environmental well-being	0.10	2.06*	0.004			
Accountability	0.04	0.81	0.001			

* $p < 0.05$, ** $p < 0.01$, ***$p < 001$

(F (1, 517) = 61.02, $p < 0.001$) and 4% of the variance in functionality trust in AI (F (1, 517) = 33.50, $p < 0.001$). As predicted, familiarity was a significant positive predictor of both trust dimensions (human-like trust: $\beta = 0.33$, $p < 0.001$; functionality trust: $\beta = 0.27$, $p < 0.001$), corroborating H2.

5.3 Perceived importance of ethics requirements of AI

RQ1 focused on the perceived importance of the seven ethics requirements of AI: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; social and environmental well-being; and accountability.

As illustrated in Fig. 2, the three ethical principles of AI considered most important were accountability ($M = 4.03$, $SD = 1.04$), transparency ($M = 3.97$, $SD = 1.02$), and privacy and data governance ($M = 3.93$, $SD = 1.09$), followed by diversity, non-discrimination and fairness ($M = 3.91$, $SD = 1.11$), technical robustness and safety ($M = 3.84$, $SD = 1.04$), human agency and oversight ($M = 3.84$, $SD = 1.04$), and societal and environmental well-being ($M = 3.76$, $SD = 1.14$). These means indicate small differences in importance ratings of the ethical requirements offered by the EU.

Table 4 Summary of hierarchical regression analyses for variables predicting functionality trust in AI ($N = 525$)

Variable	β	t	sr^2	R	R^2	ΔR^2
Step 1				0.33	0.11	0.11
Age	– 0.17	– 4.17***	0.030			
Gender	0.00	0.02	0.000			
Education	0.20	3.94***	0.027			
Income	0.12	2.23*	0.009			
Step 2				0.58	0.33	0.22
Age	– 0.17	– 4.34***	0.024			
Gender	– 0.02	– 0.56	0.000			
Education	0.10	2.19*	0.006			
Income	0.04	0.81	0.001			
Trust propensity	0.31	7.37***	0.070			
Trust in institutions	0.29	6.56***	0.055			
Step 3				0.61	0.37	0.04
Age	– 0.09	– 2.29*	0.006			
Gender	– 0.04	– 0.99	0.001			
Education	0.06	1.25	0.002			
Income	0.00	0.06	0.000			
Trust propensity	0.28	6.70***	0.055			
Trust in institutions	0.20	4.45***	0.024			
AI familiarity	0.27	5.79***	0.041			
Step 4				0.66	0.43	0.06
Age	– 0.10	– 2.61**	0.008			
Gender	– 0.01	– 0.16	0.000			
Education	0.03	0.74	0.001			
Income	– 0.03	– 0.59	0.000			
Trust propensity	0.17	4.14***	0.019			
Trust in institutions	0.20	4.48***	0.022			
AI familiarity	0.25	5.54***	0.034			
Privacy and data governance	0.01	0.14	0.000			
Human agency and oversight	0.06	1.01	0.001			
Technical robustness and safety	0.12	2.08*	0.005			
Transparency	– 0.07	– 1.20	0.002			
Diversity, non-discrimination and fairness	0.03	0.63	0.000			
Societal and environmental well-being	0.06	1.22	0.002			
Accountability	0.12	2.42*	0.007			

* $p < 0.05$, ** $p < 0.01$, ***$p < 001$

5.4 Ethics requirements and trust in AI

RQ2 asked whether people's perceptions of the importance of ethics requirements predicted trust in AI. The regression analysis results showed that after accounting for the variables entered in the first three steps, adding perceptions of the importance of the seven ethics guidelines explained an additional 2% of the variance in the variance in human-like trust in AI (F (7, 510) = 3.39, $p < 0.001$) and an additional 6% of the variance in functionality trust in AI (F (7, 510) = 7.89, $p < 0.001$).

Participants' perceived importance of one of the seven ethics requirements—societal and environmental well-being ($\beta = 0.10$, $p < 0.05$)—was a statistically significant predictor of human-like trust in AI. Two different ethics requirements, technical robustness and safety ($\beta = 0.12$, $p < 0.05$) and accountability ($\beta = 0.12$, $p < 0.05$) contributed to an increased level of functionality trust in AI.

6 Discussion

It is widely understood that AI is more than just a technology. Given its increasing importance in people's lives and the human characteristics it manifests, it is recognized as a socio-technical system that relies on human trust and must

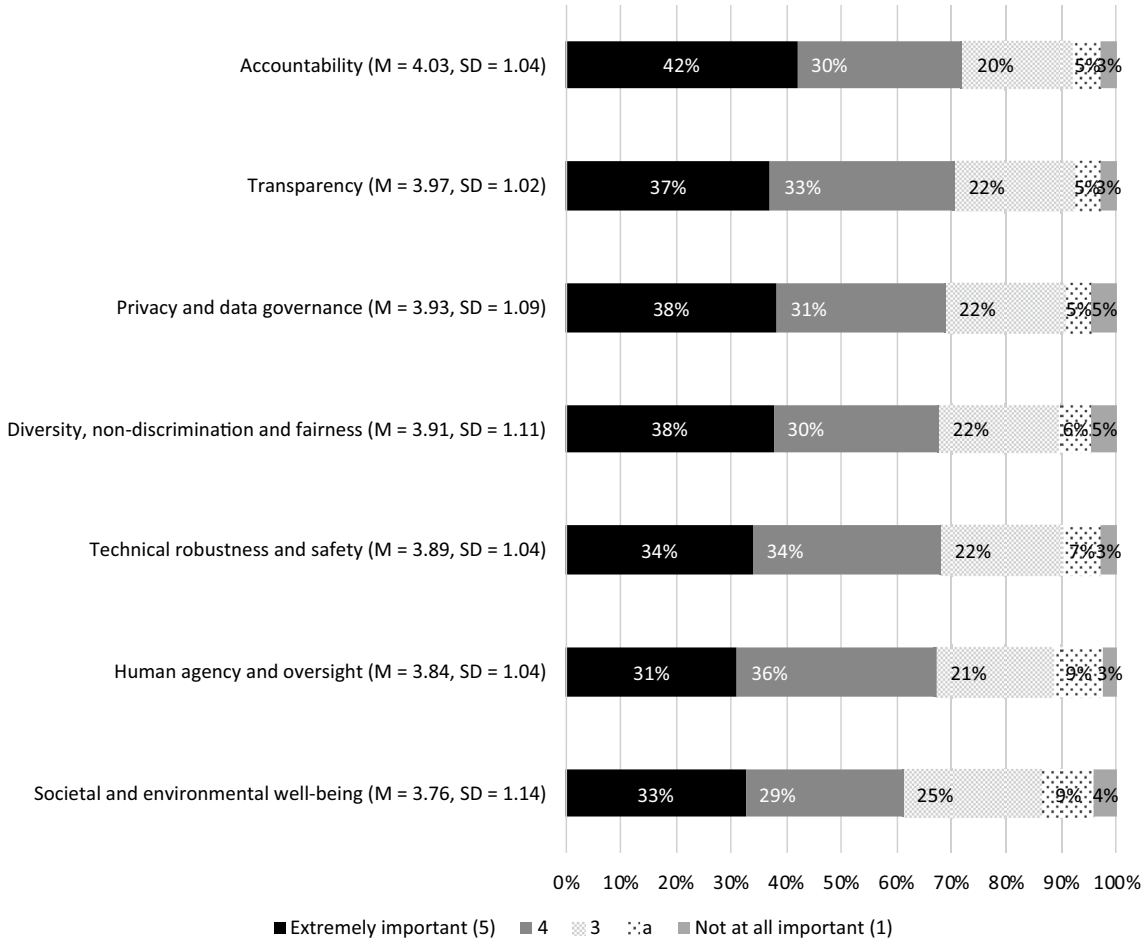

Fig. 2 Importance of ethics requirements in AI

be compliant with ethical values. Hence both trust and ethics were examined in this study. First, two dimensions emerged from our analysis—trust in the technical and functional ability of AI, and trust in the human-like characteristics of AI—which served as separate outcome variables against which demographics, levels of trust, familiarity and ethics of AI were regressed.

Age, education, and familiarity with smart consumer technologies were significant predictors of both dimensions of trust. Younger and more educated individuals indicated greater trust in AI, as did familiarity with smart consumer technologies, which was used as an indicator of experiential trust. Further, propensity for trust in other people (dispositional trust) as well as trust in institutions (institutional trust) were strongly correlated with both dimensions of trust and were significant predictors even after controlling for familiarity and demographic variables. Collectively, these findings point to an optimistic future for AI, with positive sentiments among future generations and those with more familiarity with AI consumer products. Further, trust in AI was greater than trust in other institutions, such as the government, corporations, and big tech companies, and second only to trust in other people. Overall, these findings highlight the importance of the individual, organizational, and cultural contexts in the understanding of trust in AI and its effects (Lee and See 2004). The rest of this section offers a detailed discussion on findings related to trust, followed by a discussion on the effects of ethical practices of AI on trust and the need for governance principles that conform to ethics.

6.1 Trust in AI

Results suggest that functional trust in AI is on par with trust in other people and significantly more than trust in other institutions. Furthermore, as we had expected, functional trust was higher among those who use smart home products, chatbots, conversational agents, and other consumer technologies. The positive correlation between user experience and trust augurs well for the future of AI and is in line with recent findings of positive attitudes toward AI (Helberger et al. 2020). For instance, Araujo et al. (2020) reported that decisions made by algorithms are more positively evaluated

than decisions made by human experts. This optimistic bias in perceptions of AI can be attributed to "machine heuristics" (Sundar and Kim 2019) or "algorithmic appreciation" (Logg et al. 2019). Both concepts highlight tendencies in human thinking that machines and algorithms are objective and unbiased, and therefore trustworthy. Some researchers have suggested that this positive stereotyping of technology explains our high initial trust in AI, which can erode as people accumulate experiences (Lee and See 2004).

The other dimension of trust that emerged from our analysis is human-like trust in AI, which is a combination of the attributes of human characteristics like integrity and benevolence. Combining integrity and benevolence dimensions into one dimension for AI deviates from a three-dimensional conceptualization of trust presented in previous studies. On this dimension, trust in AI was significantly lower than trust in other humans but still greater than trust in other institutions, such as governments. The gap between trust in the functionality of AI and trust in human-like characteristics can be explained in part as algorithmic aversion (Burton et al. 2020). Contrary to algorithmic appreciation (Logg et al. 2019), algorithmic aversion occurs when users realize that algorithms are imperfect, such as lacking in understanding of the human condition and as a reductive process that is cold and bereft of human emotions (Dietvorst et al. 2015; Lee 2018). These concerns about algorithms have raised awareness of the need for ethics in AI and forced a reckoning at major corporations like Google. Given the reach of AI, researchers like Timnit Gebru demand higher ethical standards to address discrimination, promote equity and contribute to overall social and environmental well-being (Simonite 2021). Though some of these demands may be beyond the reach of AI at this stage in its development, experts urge us to consider the future, referred to as the point of singularity (Arogyaswamy 2020), when machine intelligence may surpass human intelligence and offer solutions that have hitherto eluded humans. For smart machines to deliver on this promise, it is essential to develop a strong foundation of values and ethics in the AI systems that we create.

6.2 Ethics of AI

A key objective of this study was to investigate the perceived importance of the seven ethical requirements of AI advanced by the by the AI HLEG, which addresses the critical issues confronting the field. We found that accountability, transparency, and privacy and data governance had the highest ratings on importance. The focus on accountability highlights the need to establish self-governance and professional accountability mechanisms to gain and maintain trust in AI systems (Mittelstadt 2019), which can be accomplished through setting policies and training developers. Given the number of data breaches and the extensive use of personal data by corporations, it is not surprising that privacy rose to the top three in perceived importance. Along with privacy, transparency also rose to the top. Machine learning, an essential part of modern AI, has been criticized widely in the media for its "black box" approach to solutions, which may have contributed to concerns about transparency. Further, transparency is closely associated with accountability because most users are willing to share some data if corporations are more transparent of potential risks and rewards. The combination of accountability, transparency and privacy suggests that the deepest concerns in AI are centered on individual safety, which was given more importance over broader concerns such as social and environmental well-being.

Interestingly, the importance assigned to these issues by survey respondents corroborates with the issues emphasized in ethical guidelines created by professional organizations and corporations. After analyzing ethics guidelines from 22 major organizations, Hagendorff (2020) found that privacy, fairness, non-discrimination, accountability, transparency and safety were the topics that received most attention. From this analysis and our findings, it appears that both people and policies are more focused on the *do-no-harm* principle than the *do-good* principle, which holds AI to standards of improving the well-being of society and the environment. However, when the focus of AI shifts from the functionality to human-like trust in AI, the humanistic requirement (positive impact on societal and environmental well-being) was significant, offering additional justification for differentiating between functionality trust and human-like trust in AI.

Overall, our findings highlight the significance of human values and ethics in AI. To translate ethical principles into practice, a mediating governance scheme is necessary (Mittelstadt 2019). Successful governance requires a multilevel approach with interdependencies among translational/national, organizational, and individual levels. Although experts, professional organizations, and companies have announced ethical guidelines, as reviewed earlier in this paper, rules and regulatory frameworks for enforcement have not kept pace (Hagendorff 2020). Recent papers have proposed ethics-based governance through certification or accreditation programs (Roski et al. 2021) and ethics-based auditing (EBA) as governance mechanisms to help organizations realize their ethical commitments (Mökander and Floridi 2021; Mökander et al. 2021; Mökander and Axente 2021). Along with organizational-level efforts, the empowerment of citizens to learn and exercise their rights and the ability of citizens to critically evaluate AI technologies will provide a solid grounding for the development and prosperity of ethical and trustworthy AI.

Other approaches to self-governance include radical transparency, such as the TuringBox project, which requires developers to share their source code to be examined by

other developers and independent researchers (Epstein et al. 2018). While such approaches may work in academic settings, they may not have traction among corporations that employ AI for a competitive advantage. The role of competition and profit is a challenge for current ethical frameworks in AI, which are drawn from principles of beneficence, non-maleficence, autonomy and justice that are the bedrock of bioethics (Floridi and Cowls 2019). However, unlike the field of medicine, AI has no explicit "caring" obligation nor professional requirements, such as the Hippocratic oath or licensure. Further, unlike medicine, AI has no longstanding tradition of translating abstract values into concrete policies or a legal framework to impose accountability for malpractice (Mittelstadt 2019). In short, the soft approach currently in place is insufficient to enforce accountability.

As AI ethics has come under scrutiny, the application of classical theories like the Kantian categorical imperative that there are universal rights and wrongs, or the utilitarian principle of minimizing harm and maximizing good for most people have been examined. There appears to be an emerging consensus that virtue ethics, which emphasizes the moral character of the person over specific actions, is a promising avenue to create more responsible AI (Abney 2012). Virtues are dispositions to act in a certain way, and humans involved in the development of AI can be trained to develop virtues and integrate ethics in all aspects of design and development. But passing on virtue ethics to an autonomous agent or AI system is a formidable task because the understandings of context, intentionality, complex thoughts, and consequences of moral actions remain elusive in AI (Allen and Wallach 2012). Despite this limitation, virtue ethics is a promising approach through which humans can integrate ethics in all their interactions with AI. In summary, a combination of bottom-up virtue ethics and top-down regulatory frameworks that hold organizations more accountable appears to be a pragmatic way forward for developing ethical AI.

6.3 Limitations and future directions

Our findings offer empirical evidence of the correlation between ethical principles and trust in AI, which must be evaluated keeping in mind some of the limitations of this study. First, we focused on AI technology only in consumer products (i.e., smart technologies), which is limited in scope. Future research must explore ethical requirements of AI in specific domains such as health, criminal justice, or autonomous vehicles in which trust is critical for success. In future work, a better understanding of fairness, transparency, accountability, safety and robustness within each domain will be critical for widespread acceptance of these technologies. Second, the survey offers a snapshot of the public's level of trust and attitudes toward ethics at the current moment. However, researchers have pointed out that trust is a dynamic construct that evolves over time that can erode quickly after an adverse event (Hoff & Bashir, 2015), which underscores the need for a longitudinal approach to the study of trust. Third, the findings from this study are correlational, and we cannot assume causation between ethics of AI and trust in AI. Lastly, as with any online panel, our study respondents may not be truly representative of the U.S. population. The fact that all study participants had access to the internet skews the sample toward those with greater access to and experience with various forms of information technology. Future studies should examine more representative samples with varying levels of experience with technology.

7 Conclusion

During a time of increased misgivings in human-established institutions, trust in AI remains relatively high. Instead of a laissez-faire approach, this is the time to incorporate ethics and values in AI systems and nurture these sensibilities among technologists through training in virtue ethics. As we develop intelligent artifacts that may someday surpass human intelligence, it behooves us to build internal mechanisms that do not perpetuate human biases. It is also important for these systems to root out bad data and malevolent actors. There is an equally pressing imperative to build AI systems with an ethical compass that elevates human life and creates a society in which humans and machines interact in harmony. Human-AI interaction must strive for higher goals that go beyond productivity and profitability to foster understanding and compassion. Furthermore, AI systems must promote compassion and understanding between humans. For example, a polite conversational agent with a calm voice could perpetuate polite behaviors in humans, especially as the use of such agents becomes pervasive in modern society. Though humans have struggled to achieve these goals, the optimistic ambition is that with assistance from AI, we may get closer to these ideals, which is at the heart of trustworthy AI.

Appendix: Survey questionnaires, scales, and reliability coefficients

Variable	Survey items	Scale	Reliability
Trust propensity	I usually trust people until they give me a reason not to trust them	1 (strongly disagree) – 5 (strongly agree)	$\alpha = 0.85$
	I generally give people the benefit of the doubt when I first meet them		
	My typical approach is to trust new acquaintances until they prove I should not trust them		
Trust in intuitions	To what extent do you trust the following institutions? [Federal government]	1 (do not trust) – 5 (highly trust)	$\alpha = 0.90$
	To what extent do you trust the following institutions? [Corporations]		
	To what extent do you trust the following institutions? [Big technology companies]		
Familiarity with AI technologies	Here are some examples of smart technology that we encounter every day, which uses AI. How often do you use these technologies? [Smart home devices (e.g., Google Nest, Ring, Blink)]	1 (never) – 5 (very frequently)	$\alpha = 0.88$
	How often do you use these technologies? [Smart speakers (e.g., Amazon Echo, Google Home, Apple Homepod, Sonos)]		
	How often do you use these technologies? [Virtual assistants (e.g., Siri, Alexa, Cortana)]		
	How often do you use these technologies? [Wearable devices (e.g., Fitbit, Apple Watch)]		
Importance of ethics principles	How **important** are these **values** in the design of AI and smart technologies that interact with us? [**Privacy and data governance:** Competent authorities who implement legal frameworks and guidelines for testing and certification of AI-enabled products and services.]	1 (not at all important) – 5 (extremely important agree)	
	How **important** are these **values** in the design of AI and smart technologies that interact with us? [**Human agency and oversight**: Human oversight and control throughout the lifecycle of AI products.]		
	How **important** are these **values** in the design of AI and smart technologies that interact with us? [**Technical robustness and safety**: Systems are developed in a responsible manner with proper consideration of risks.]		
	How **important** are these **values** in the design of AI and smart technologies that interact with us? [**Transparency:** Transparency requirements that reduce the opacity of systems.]		
	How **important** are these **values** in the design of AI and smart technologies that interact with us? [**Diversity, non-discrimination and fairness**: The application of rules designed to protect fundamental human rights, such as equality.]		
	How **important** are these **values** in the design of AI and smart technologies that interact with us? [**Societal and environmental well-being:** AI systems that conform to the best standards of sustainability and address like issues climate change and environmental justice.]		
	How **important** are these **values** in the design of AI and smart technologies that interact with us? [**Accountability**: AI at any step is accountable for considering the system's impact in the world.]		
Human-like trust in AI	Smart technologies care about our well-being. (Benevolence)	1 (strongly disagree) – 5 (strongly agree)	$\alpha = 0.92$
	Smart technologies are sincerely concerned about addressing the problems of human users. (Benevolence)		
	Smart technologies try to be helpful and do not operate out of selfish interest. (Benevolence)		
	Smart technologies are truthful in their dealings. (Integrity)		
	Smart technologies keep their commitments and deliver on their promises. (Integrity)		
	Smart technologies are honest and do not abuse the information and advantage they have over their users. (Integrity)		
Functionality trust in AI	Smart technologies work well. (Competence)	1 (strongly disagree) – 5 (strongly agree)	$\alpha = 0.91$
	Smart technologies have the features necessary to complete key tasks. (Competence)		
	Smart technologies are competent in their area of expertise. (Competence)		
	Smart technologies are reliable. (Competence)		
	Smart technologies are dependable. (Competence)		

Data availability The dataset analyzed in the current study is available from the corresponding author on request.

Declarations

Conflict of interest The authors declare that there is no conflict of interest. This project was funded in part by a National Association of Broadcasters Pilot Grant.

References

Abney K (2012) Robotics, ethical theory, and metaethics: A guide for the perplexed. In: Lin P, Abney K, Bekey GA (eds) Robot ethics: the ethical and social implications of robotics, First MIT Press, paperback. The MIT Press, Cambridge, Massachusetts London, England, pp 35–54

Alarcon GM, Lyons JB, Christensen JC et al (2018) The effect of propensity to trust and perceptions of trustworthiness on trust behaviors in dyads. Behav Res Methods 50:1906–1920. https://doi.org/10.3758/s13428-017-0959-6

Allen C, Wallach W (2012) Moral machines: Contradiction in terms of abdication of human responsibility? In: Lin P, Abney K, Bekey GA (eds) Robot ethics: the ethical and social implications of robotics, First MIT Press, paperback. The MIT Press, Cambridge, Massachusetts London, England, pp 55–68

Araujo T, Helberger N, Kruikemeier S, de Vreese CH (2020) In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI Soc 35:611–623. https://doi.org/10.1007/s00146-019-00931-w

Arogyaswamy B (2020) Big tech and societal sustainability: an ethical framework. AI Soc 35:829–840. https://doi.org/10.1007/s00146-020-00956-6

Borgesius FJ (2018) Discrimination, artificial intelligence, and algorithmic. Directorate General of Democracy, Council of Europe, Strasbourg

Burton JW, Stein M-K, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. J Behav Decis Mak 33:220–239. https://doi.org/10.1002/bdm.2155

Calhoun CS, Bobko P, Gallimore JJ, Lyons JB (2019) Linking precursors of interpersonal trust to human-automation trust: An expanded typology and exploratory experiment. J Trust Res 9:28–46. https://doi.org/10.1080/21515581.2019.1579730

Chatila R, Havens JC (2019) The IEEE global initiative on ethics of autonomous and intelligent systems. In: Aldinhas Ferreira MI, Silva Sequeira J, Singh Virk G et al (eds) Robotics and Well-Being. Springer International Publishing, Cham, pp 11–16

Chen SC, Dhillon GS (2003) Interpreting Dimensions of Consumer Trust in E-Commerce. Inf Technol Manag 4:303–318

Choung H, David P, Ross A (2022) Trust in AI and its role in the acceptance of AI technologies. Int J Hum-Comput Interact. https://doi.org/10.1080/10447318.2022.2050543

Colquitt JA, Scott BA, LePine JA (2007) Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. J Appl Psychol 92:909–927. https://doi.org/10.1037/0021-9010.92.4.909

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 144:114–126. https://doi.org/10.1037/xge0000033

Edelman (2021) Edelman trust barometer 2021

Epstein Z, Payne BH, Shen JH, et al (2018) TuringBox: An experimental platform for the evaluation of AI systems. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden, pp 5826–5828

Floridi L, Cowls J (2019) A unified framework of five principles for AI in society. Harv Data Sci Rev. https://doi.org/10.1162/99608f92.8cd550d1

Floridi L, Cowls J, Beltrametti M et al (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach 28:689–707. https://doi.org/10.1007/s11023-018-9482-5

Frazier ML, Johnson PD, Fainshmidt S (2013) Development and validation of a propensity to trust scale. J Trust Res 3:76–97. https://doi.org/10.1080/21515581.2013.820026

Fulmer A, Dirks K (2018) Multilevel trust: a theoretical and practical imperative. J Trust Res 8:137–141. https://doi.org/10.1080/21515581.2018.1531657

Gefen D (2000) E-commerce: the role of familiarity and trust. Omega 28:725–737. https://doi.org/10.1016/S0305-0483(00)00021-9

Gefen D, Karahanna E, Straub DW (2003) Trust and TAM in online shopping: an integrated model. MIS Q 27:51–90. https://doi.org/10.2307/30036519

Gillath O, Ai T, Branicky MS, et al (2021) Attachment and trust in artificial intelligence. Comput Hum Behav 10

Gulati R (1995) Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances. Acad Manag J 38:85–112. https://doi.org/10.2307/256729

Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Minds Mach 30:99–120. https://doi.org/10.1007/s11023-020-09517-8

Hancock PA, Billings DR, Schaefer KE et al (2011) A meta-analysis of factors affecting trust in human-robot interaction. Hum Fact 53:517–527. https://doi.org/10.1177/0018720811417254

Helberger N, Araujo T, de Vreese CH (2020) Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. Comput Law Secur Rev 39:105456. https://doi.org/10.1016/j.clsr.2020.105456

Hleg AI (2019) Ethics guidelines for trustworthy AI. European Commission, Brussels

Hoff KA, Bashir M (2015) Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors J Hum Factors Ergon Soc 57:407–434. https://doi.org/10.1177/0018720814547570

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1:389–399. https://doi.org/10.1038/s42256-019-0088-2

Lankton N, McKnight DH, Tripp J (2015) Technology, humanness, and trust: Rethinking trust in technology. J Assoc Inf Syst 16:880–918. https://doi.org/10.17705/1jais.00411

Lee MK (2018) Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. Big Data Soc 5:205395171875668. https://doi.org/10.1177/2053951718756684

Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. Hum Factors 46:50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. Org Behav Hum Decis Process 151:90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Madhavan P, Wiegmann DA (2007) Effects of information source, pedigree, and reliability on operator interaction with decision support systems. Hum Fact J Hum Fact Ergon Soc 49:773–785. https://doi.org/10.1518/001872007X230154

Mayer RC, Davis JH, Schoorman FD (1995) An integrative model of organizational trust: past, present, and future. Acad Manage Rev 20:709–734

Mcknight DH, Carter M, Thatcher JB, Clay PF (2011) Trust in a specific technology: an investigation of its components and measures.

ACM Trans Manag Inf Syst 2:1–25. https://doi.org/10.1145/1985347.1985353
Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1:501–507. https://doi.org/10.1038/s42256-019-0114-4
Mökander J, Axente M (2021) Ethics-based auditing of automated decision-making systems: intervention points and policy implications. AI Soc. https://doi.org/10.1007/s00146-021-01286-x
Mökander J, Floridi L (2021) Ethics-based auditing to develop trustworthy AI. Minds Mach 31:323–327. https://doi.org/10.1007/s11023-021-09557-8
Mökander J, Morley J, Taddeo M, Floridi L (2021) Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. Sci Eng Ethics 27:44. https://doi.org/10.1007/s11948-021-00319-4
OECD (2019) Artificial intelligence in society. OECD Publishing, Paris
Roski J, Maier EJ, Vigilante K et al (2021) Enhancing trust in AI through industry self-governance. J Am Med Inform Assoc 28:1582–1590. https://doi.org/10.1093/jamia/ocab065
Rotenberg KJ (2019) The psychology of interpersonal trust: theory and research. Routledge, Abingdon, Oxon, New York
Rousseau DM, Sitkin SB, Burt RS, Camerer C (1998) Not so different after all: a cross-discipline view of trust. Acad Manag Rev 23:393–404. https://doi.org/10.5465/amr.1998.926617
Schoorman FD, Mayer RC, Davis JH (2007) An integrative model of organizational trust: Past, present, and future. Acad Manag Rev 32:344–354. https://doi.org/10.5465/amr.2007.24348410
Shin D (2021) The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. Int J Hum-Comput Stud 146:102551. https://doi.org/10.1016/j.ijhcs.2020.102551
Shin D, Park YJ (2019) Role of fairness, accountability, and transparency in algorithmic affordance. Comput Hum Behav 98:277–284. https://doi.org/10.1016/j.chb.2019.04.019
Simonite T (2021) What Really Happened When Google Ousted Timnit Gebru. Wired
Sundar SS, Kim J (2019) Machine heuristic: when we trust computers more than humans with our personal information. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19. ACM Press, Glasgow, Scotland Uk, pp 1–9
Thiebes S, Lins S, Sunyaev A (2021) Trustworthy Artificial Intelligence Electron Mark 31:447–464. https://doi.org/10.1007/s12525-020-00441-4
Torresen J (2018) A review of future and ethical perspectives of robotics and AI. Front Robot AI 4:75. https://doi.org/10.3389/frobt.2017.00075
Wu K, Zhao Y, Zhu Q et al (2011) A meta-analysis of the impact of trust on technology acceptance model: investigation of moderating influence of subject and context type. Int J Inf Manag 31:572–581. https://doi.org/10.1016/j.ijinfomgt.2011.03.004

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.