A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions

Steven Lockey
University of Queensland
s.lockey@uq.edu.au

Nicole Gillespie
University of Queensland
n.gillespie@business.uq.edu.au

Daniel Holm
University of Queensland
d.holm@business.uq.edu.au

Ida Asadi Someh
University of Queensland
i.asadi@business.uq.edu.au

Abstract

Artificial Intelligence (AI) can benefit society, but it is also fraught with risks. Societal adoption of AI is recognized to depend on stakeholder trust in AI, yet the literature on trust in AI is fragmented, and little is known about the vulnerabilities faced by different stakeholders, making it is difficult to draw on this evidence-base to inform practice and policy. We undertake a literature review to take stock of what is known about the antecedents of trust in AI, and organize our findings around five trust challenges unique to or exacerbated by AI. Further, we develop a concept matrix identifying the key vulnerabilities to stakeholders raised by each of the challenges, and propose a multi-stakeholder approach to future research.

1. Introduction

Artificial Intelligence (AI) is an increasingly ubiquitous aspect of modern life that has had a transformative impact on how we live and work [1]. However, despite holding much promise AI has been implicated in high profile breaches of trust and ethical standards and concerns have been raised over the use of AI in initiatives and technologies that could be inimical to society. For example, AI underpins lethal autonomous weapons, is central to mass surveillance, and is subject to racial bias in healthcare.

Trust is vital for AI's continued social license. The European Commission's AI High-Level Expert Group (AI HLEG) highlight that if AI systems do not prove to be worthy of trust, their widespread acceptance and adoption will be hindered, and the vast potential societal and economic benefits will remain unrealized [2]. While trust has been shown to be important for the adoption of a range of technologies [3], AI creates an array of qualitatively different trust challenges compared to more traditional information technologies [4]. In response, the AI HLEG provided a set of guidelines for the development, deployment and use of trustworthy AI [2]. These guidelines are just one of many [5].

Research shows that trust is an important predictor of the willingness to adopt a range of AI systems, from product recommendation agents [e.g., 6, 7] and AI-enabled banking [e.g., 8] to autonomous vehicles (AVs) [e.g., 9, 10]. Given the central role of trust, there is a strong practical need to understand what influences and facilitates trust in AI, with multiple recent calls for research from policymakers [2, 11], industry [12] and scholars [e.g., 13, 14].

Yet we are only at an early stage of understanding the antecedents of trust in AI systems. A recent review of the empirical literature suggests that AI representation plays an important role in the development of trust [15] and differentially impacts trust over time; for robotic AI, trust tends to start low and increase over time, but for virtual and embedded AI the opposite commonly occurs. However, it is difficult however to isolate the antecedents of trust in this work, as trust was equated with affect [e.g. 16] attraction to [e.g. 17] and general perceptions of AI [e.g. 18]. Previous meta-analyses have examined the antecedents to trust in specific applications of AI, such as human-robot interaction [19] and automation [20], but have not taken into account human trust in AI more broadly.

In this review, we take stock of the scholarly literature over the past two decades to examine the antecedents of trust in AI systems. Our review differs to prior work in four ways: 1) our organization of the literature around five trust challenges that are unique to, or exacerbated by, the inherent characteristics of AI; 2) our focus on articles that operationalize trust in line with established definitions; 3) a focus on trust in all forms of AI; and 4) the integration of conceptual and empirical scholarship.

We contribute to the literature on trust in AI in three ways. First by synthesizing the fragmented and interdisciplinary literatures to meaningfully take stock of what we know about the antecedents of trust in AI. Second, by developing a concept matrix identifying the key vulnerabilities for stakeholders raised by each of the five AI trust challenges. Third, by drawing on this matrix to identify omissions in current

URI: https://hdl.handle.net/10125/71284
978-0-9981331-4-0

understanding and promising directions for future research.

2. Defining AI and Trust

2.1. Conceptualizing AI

We adopt the OECD's [21] definition of AI, as recently recommended by AI experts [22]: "*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*".

Most notable advances in AI are driven by machine learning [23], a subset of AI and can be defined as a "machine's ability to keep improving its performance without humans having to explain exactly how to accomplish all the tasks it's given" [34]. A further subset of machine learning is deep learning, which is a specialized class of machine learning that is built on artificial neural networks [25]. Advances in machine learning and the shift from rule-based to algorithmic learning exponentially increases the power and functionality of these systems, enabling more accurate results than previous iterations. However, they also change the nature of how IT artifacts are designed and work [26], their capacity for autonomous functioning, creating risks, challenges and uncertainties [27] not inherent in traditional technologies. Trust matters most under conditions of risk and uncertainty [28, 29].

2.2. Conceptualizing trust

We adapt popular, cross-disciplinary definitions [30, 31] to define trust as a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviour of another entity (e.g. an AI system).

The two defining components of trust are the intention to accept *vulnerability* based on *positive expectations.* In positioning their stance on trust in IT artifacts, McKnight et al. [32, p. 3] note: "trust situations arise when one has to make oneself vulnerable by relying on another person or object". Trust is only relevant under conditions of risk and uncertainty, where misplaced trust results in loss or harm [32]. Examples include relying on an autonomous vehicle to drive safely, or on the decision of an AI system to be accurate and unbiased. Vulnerability is central to trust and captures the 'leap of faith' required to engage with entities under conditions of risk and uncertainty.

A foundational tenet of trust theory is that this willingness to be vulnerable should be based on 'good reasons' [33]. 'Trusting' without good reasons (or positive expectations) is not trust at all; it amounts to hope or blind faith. Positive expectations of AI systems can be based on system-oriented assessments of functionality, reliability and predictability, and helpfulness [32]. Hence, there must be some expected utility or value to accept vulnerability to an AI system – that is, positive expectations that the system will be useful, reliable and operate as intended.

Trust theory and research highlights the importance of understanding the trustor (i.e. who is doing the trusting), the referent of trust (i.e. what or whom are they trusting in), and the nature of trusting (i.e. what are the risks, vulnerabilities or dependence in the trusting act) [34, 35, 36]. Understanding the trustor (i.e. the stakeholder) is particularly important in the context of AI, as it will influence the nature of the risks and vulnerabilities inherent in trusting an AI system, and hence the salient cues and antecedents that influence trust. For example, domain experts are likely to pay attention to different trust cues than those that impact end users or customers.

3. Methodology

We conducted an interdisciplinary literature review using the Web of Science and EBSCO Business Source Complete databases, searching for the terms "*trust*" AND "Artificial Intelligence" OR "Machine Learning" OR "Deep Learning". Peer-reviewed journal articles, conference and symposia papers and proceedings, and book chapters published since 2000 were included in our review. We further examined the reference lists of recent review articles on trust in AI, robots and automation [e.g. 15, 19] and highly cited papers [e.g. 13] to identify additional articles that met our inclusion criteria.

We excluded articles that did not address antecedents of trust in AI, either conceptually or empirically, and did not meet a commonly accepted definition or conceptualization of trust (e.g., where trust was conflated with distinct constructs, such as emotion or attraction). Reasons for exclusion included: a focus on computational trust, discussion of trusts in the financial/legal sense (e.g., trust fund) or healthcare (e.g., an NHS trust), articles in which trust was peripheral rather than central to the article, or in empirical papers that mention trust but did not measure it. After this screening process, our search produced 102 relevant articles.

Our review comprised more empirical (57%) than conceptual (43%) articles. Most empirical papers were experimental (47/58 papers), and only one paper used a mixed-method design. 71% of papers were published in 2016 or later, and the earliest article in our review was published in 2005. Articles reflected a diversity of fields, including information systems, computer science, ergonomics, business and economics, psychology, medicine, and law.

4. Literature Review: AI Trust Challenges

We organize our review by focusing on concepts related to five central AI trust challenges: 1) transparency and explainability, 2) accuracy and reliability, 3) automation, 4) anthropomorphism, and 5) mass data extraction. These five trust challenges capture the large majority of articles identified by our review. This approach positions our paper as an *organizing review* [37]. For each concept, we first explain the trust challenge, before synthesizing the relevant literature.

4.1. Transparency and Explainability

AI is often considered a 'black box' [38]. Advanced algorithmic learning methods (such as deep learning) are inherently not transparent or explainable. The antidote to this black box is creating AI that can explain itself, where decisions and predictions are made transparently. However, there is a tension between accuracy and explainability, in that models that perform best tend to be the least transparent and explainable, while the ones most able to provide the clearest explanations are the least accurate [39]. There is an entire field of research dedicated to making AI more explainable and transparent, with the central aim of improving user trust [38].

Many articles in our review theorize or empirically demonstrate that transparency and explainability of AI applications facilitate trust. In healthcare, scholars argue that interpretable models that are explainable and transparent are necessary to enable clinicians to understand and trust in the outcomes of clinical support systems [40, 41]. However, full transparency may be difficult to achieve in practice. Instead, different levels of transparency can be used based on factors such as level of risk and the ability of the clinician to evaluate the decision [41].

Explanations are argued to play a key role in facilitating trust in AI systems [42], particularly when the user lacks previous experience with the system. Researchers propose that system transparency is a key mitigator of user overtrust, that is trusting AI more than is warranted by its capabilities [43, 44]. However, explanations may actually *cause* overtrust [45] and can be manipulative [46]. The seminal 'Copy Machine' study [47] showed that providing an explanation, even without a legitimate reason, was effective in promoting compliance. This is particularly problematic when the audience of the explanation (e.g. an end user) diverges from its beneficiary (e.g. a deploying organization; [46]). System explanations are problematic when produced alongside incorrect results, particularly when they seem plausible [45].

Empirical research demonstrates the positive impact of AI transparency and explainability on trust [e.g. 48, 49, 50]. Experimental research undertaken in military settings indicates that when human operators and AI agents collaborate, increased transparency enhances trust [48,49]. Explanations have been shown to increase trust in the results of a product release planning tool [51].

However, research further suggests that the relationship between transparency and trust is not straightforward. For example, in the context of students interacting with an AI assessment grading tool, providing procedural transparency about the fairness of the algorithm was found to buffer the negative impact of expectation violation on trust [52]. However, providing more transparency related to the outcome (how the raw grade was calculated) did not enhance trust, indicating that the type and amount of transparency matters.

4.2. Accuracy and Reliability

A key trust challenge relates to the accuracy of AI systems, as inaccurate outcomes can lead to bias, inequality and harm. AI systems can be configured to optimize a variety of accuracy metrics, and may have a high rate of accuracy for certain predictions (e.g. outcomes for white men), but not others (e.g. outcomes for minority groups) [53]. A study of automated facial analysis algorithms demonstrated this problem; there were significantly more misclassifications of darker-skinned females than lighter-skinned males [54]. Hence, relying on accuracy metrics alone may not be sufficient to garner trust in AI applications; the fairness of the system is also relevant [53].

Several experiments show that as the reliance, accuracy or performance of AI systems decreases, so does user trust [55, 56]. The timing of a reliability failure also matters. Unreliable performance early in one's experience with a system may cause more significant trust break down than failure later in an interaction [57]. Moreover, even if an AI agent is accurate, users may not trust it [58]: they also need to *perceive* that it is accurate. For example, teams engaged in a large, street-based game were regularly mistrustful of the (entirely accurate) information provided by automated advice, and often chose to ignore it, despite being told that following the information was vital for them to progress in the game [58].

However, other research suggests that even though inaccurate agent behaviour negatively impacts perceived trustworthiness, this does not necessarily translate into reduced compliance: users may still follow instructions from an AI system they believe is untrustworthy [59]. Taken together, while most research indicates a positive influence of accuracy on trust, the relationship is not straightforward and warrants further research.

4.3. Automation versus augmentation

Automation enables machines to complete tasks without direct human involvement [60]. Normative prescriptions tend to advise organizations to prioritize augmentation – human collaboration with machines to perform a task - over automation. Yet there is an argument that such neat delineation is not realistic and that an automation-augmentation paradox exists [60]. As an example, domain experts may work with an AI system to determine and codify appropriate variables (augmentation), and the system may then be automated based on these criteria. However, if conditions change over time, a further stage of augmentation will be necessary. This brings into question the role of the domain expert and the potential for their role in the augmentation process to ultimately lead to the automation of their own work.

The impact of automated AI on trust in high-risk contexts has been conceptually discussed. In healthcare, there are concerns that AI may disrupt the bond of trust between doctors and patients [61], and patients may be more skeptical of automated advice than advice provided by a doctor [62]. A ‘doctor-in-the-loop’ approach, in which the doctor both provides tacit knowledge to AI systems and is the final authority on decisions proposed by the AI systems, has been proposed to address these concerns [63]. This ‘augmentation over automation’ approach has received empirical support. A suite of experiments found a reluctance to use medical care delivered by AI providers, except when the AI was used to support the human provider’s decision, rather than replacing the human [64]. This ‘human-in-the-loop’ approach has also been proposed for AI in financial services [65].

Adaptive automation, where automation is not fixed at the design stage but rather adapts to the situation, increased trust in a robot during a collaborative task to a greater extent than when there was either no automation or static automation.

A concern related to automated AI is the potential for deskilling if domain experts over-rely on automated systems [67, 68]. One study found financial investors trust fully automated artificial advisors more than human advisors [69]. However other research indicates that AI over-reliance on AI systems tends to be experienced by novices; experts are generally less willing to trust AI systems [70, 71].

4.4. Anthropomorphism and embodiment

Anthropomorphism involves the inclusion of human-like characteristics into an AI’s design. It has been theorized that the more human-like an AI agent is, the more likely humans are to trust and accept it [72]. However, there are concerns that over-anthropomorphism may lead to overestimation of the AI’s capabilities, potentially putting the stakeholder at risk [73], damaging trust [74], and leading to a host of ethical and psychological concerns, including manipulation [75].

Empirical findings broadly support the proposition that anthropomorphism increases trust in AI. This has been shown in the context of autonomous vehicles [72,76], with people demonstrating more trust in AVs with human features than without [72], as well as in the context of virtual agents [e.g. 9, 77].

Research into the buffering impact of virtual agent human-likeness on decreasing reliability found that although anthropomorphism decreased initial expectations, it increased trust resilience. When performance deteriorated, decreases in trust were more pronounced in a machine-like agent than an anthropomorphic agent. Embodiment of virtual agents (i.e. having a physical form) also increases user trust in the agent, primarily through perceptions of its social presence [9, 77, 78]. Research also indicates that augmented reality and 3D agents were perceived as more trustworthy than those in traditional 2D interfaces [79].

However, not all empirical work suggests that anthropomorphism leads to stronger perceptions of trust. For example, a study investigating the anthropomorphism of a care robot found that a highly human-like robot was perceived as less trustworthy and empathetic than a more machine-like robot [62]. Further research is required to understand when and how AI anthropomorphism enhances trust, and what moderates this relationship.

4.5. Mass Data Extraction

AI systems, particularly advanced algorithmic learning systems, require the extraction and processing of large amounts of data to function, making them qualitatively different from traditional IT artifacts [81]. Data extraction is fundamentally different from previous forms of market exchange, as it connotes a one-way taking of data rather than a consensual or reciprocal process [82].

The trust challenge around data extraction is primarily around issues of privacy. For end users, loss of privacy and inappropriate sharing of information is a concern, and can result in reduced self-determination. These vulnerabilities can scale to the societal level to the point where expectations of privacy as a societal norm may be lost. Indeed, Facebook CEO Mark Zuckerberg has explicitly stated that privacy is no longer a social norm [83]. This proposition is clearly contentious, as privacy is considered and codified as a fundamental human right in several democracies, and people usually express that they value privacy, even if they do not always demonstrate this proposition in their behavior [84].

Some jurisdictions have taken regulatory approaches to tackling concerns about big data extraction, with the European Commission's General Data Protection Regulation (GDPR) aiming to give European residents control over their personal data through requirement of 'Privacy by Design' [85]. While this type of legislation may reduce privacy-related vulnerabilities of end-users and society, it introduces a new set of vulnerabilities for domain experts, who are responsible for ensuring data privacy and accountable for appropriate data use under threat of large fines

Research on data extraction and the privacy concerns that underpin it has been primarily conceptual. Scholars note big data extraction is an ethical dilemma in the development and use of AI-enabled medical systems [62, 86], virtual agents [87] and smart cities [88]. One solution to ensure citizen privacy and promote trust is creating an environment in which data analysis can occur without allowing organizations to extract the data [88].

The limited empirical work in this area has focused on the interaction between privacy and trust. For example, when people have few privacy concerns related to autonomous vehicles collecting passenger location information and being used as a conduit for surveillance, they were more likely to trust in the autonomous vehicle [89].

Interestingly, a study of virtual agent embodiment found that participants were more willing to share private data with an AI agent and more confident that the agent would respect their privacy when it could move around naturally and speak compared with a static agent that could speak [77].

4.6. The Role of Governance in Addressing AI Trust Challenges

In addition to the five trust challenges, our review identified two broad, generic mechanisms for overcoming these trust challenges: familiarity and governance. Empirical studies indicate that familiarity and experience engaging with AI systems facilitates trust [90, 91]. Conceptual work argues that governance – in the form of appropriate controls to ensure trustworthy AI development and deployment - is a necessary condition for trust in AI [e.g. 92, 93]. A recent national survey identified beliefs about the adequacy of AI regulation and governance to be the strongest predictor of trust in AI systems [94]. It may be more important and efficient to make AI systems verifiably trustworthy via appropriate governance rather than seek explanations for specific outcomes [45]. Governance that encourages collaboration among key stakeholders, supports the recognition and removal of bias, and clarifies the appropriate control over and use of personal information has been proposed to enhance trust [92]. However, this work further notes that AI development remains largely unregulated to date [95], despite public expectation of AI regulation [94; 96].

5. Discussion and Future Directions

Our review demonstrates that research on the antecedents of trust in AI can largely be organized around five key trust challenges that are unique to, or exacerbated by, the inherent features of AI. Each of these trust challenges raises a set of vulnerabilities or risks for stakeholders of AI systems. In Table 1, we present a concept matrix mapping the key vulnerabilities associated with each of the five trust challenges for three AI stakeholder groups – domain experts, end users, and society. These stakeholders are each central to the acceptance and uptake of AI systems.

As shown in Table 1, the use of AI systems open up (potential or actual) risks and vulnerabilities for each of these stakeholders, making trust a highly salient and pertinent concern. Our concept matrix shows that the vulnerabilities experienced in relation to an AI system depend on the stakeholders' role which determines how they interact with, are responsible for, or are impacted by the AI systems. In the next section, we discuss the key vulnerabilities domain experts, end users and society more broadly experience in relation to each AI trust challenge, and how these differ across these stakeholder groups.

Table 1: Concept matrix of the five AI trust challenges and the respective vulnerabilities each creates for stakeholders

AI trust challenge	Stakeholder vulnerabilities		
	Domain expert	End user	Society
1. Transparency and explainability	• Ability to know and explain AI output, and provide human oversight • Manipulation from erroneous explanations	• Ability to understand how decisions affecting them are made • Ability to provide meaningful consent and exercise agency	• Knowledge asymmetries • Power imbalance and centralization • Scaled disempowerment
2. Accuracy and reliability	• Accountability for accuracy and fairness of AI output • Reputational and legal risk	• Inaccurate / harmful outcomes • Unfair / discriminatory treatment	• Entrenched bias / inequality • Scaled harmed to select populations
3. Automation	• Professional over-reliance and deskilling • Loss of expert oversight • Loss of professional identity • Loss of work	• Loss of dignity (humans as data points; de-contextualization) • Loss of human engagement • Over-reliance and deskilling	• Scaled deskilling • Reduced human connection • Scaled technological unemployment • Cascading AI failures
4. Anthropomorphism and embodiment	• Professional over-reliance • Psychological wellbeing	• Manipulation through identification • Over-reliance and over-sharing	• Manipulation through identification • Human connection and identity
5. Mass data extraction	• Accountability for privacy and use of data • Reputational and legal risk	• Personal data capture and loss of privacy • Inappropriate re-identification and use of personal data • Loss of control	• Inappropriate use of citizen data • Mass surveillance • Loss of societal right to privacy • Power imbalance & societal disempowerment

Domain experts. Domain experts in deploying organizations are those with the expert knowledge and experience in the field of application of the AI system. For example, doctors in relation to AI-enabled medical diagnosis applications. Domain expert knowledge can be used to create codified information used to train AI systems, meaning they have a role in system input. Domain experts also work with system outputs, as they use and interface with AI systems for service delivery.

Key vulnerabilities faced by domain experts relate to professional knowledge, skills, identity, and reputation. For example, research suggests that automation through AI may lead to deskilling [67, 68]. A related vulnerability stemming from the AI explanability challenge is the ability of the domain expert to understand the AI system and be able to explain and justify decisions to other stakeholders, particularly when AI system outputs are used in service delivery (e.g. clinical decision making systems). Anthropomorphism may further threaten the professional identity of domain experts and cause over-reliance on human-like agents. The reputational damage and legal risks from inaccurate or unfair results, or inappropriate data use, sharing or privacy breach, place a further burden on accountable domain experts.

End users. End users are those directly influenced by the output or decisions made by the AI system. They are vulnerable to any problems, inaccuracies or biases within the system. More broadly, end users face vulnerabilities around understanding how AI-based decisions are made, which can lead to diminished ability to provide meaningful consent, identify unfair or unethical impact, and exercise agency. Using the context of AI in personal insurance as an example, companies purportedly draw on thousands of data points to judge the risk of someone making a motor insurance claim, including whether they drink tap or bottled water [97]. Understanding exactly *how* such a decision was made is impossible for an average customer, and highlights vulnerabilities around explainability, data capture and loss of privacy related to data extracted without consent. Further, AI can be used to ‘nudge’ customer behavior in a way that is manipulative and intrusive [97]. Concerns have been raised that the combination of these vulnerabilities may lead to the loss of human dignity, and lack of consideration of personal circumstances, effectively reducing humans to a series of data points. This is particularly problematic for underrepresented, marginalised users.

Society. The focus here is on vulnerabilities that impact society as a whole, and this stakeholder group includes regulators. Vulnerabilities at the societal level include knowledge asymmetry, power centralization and the potential for cascading AI failures. For instance, knowledge asymmetry between big tech companies, policymakers and citizens may result in a continuous cycle of outdated or ineffective regulation [98]. Internet giants at the forefront of AI development and mass data extraction activities have already amassed a unique concentration of power [99]. The scaled use of inaccurate, biased or privacy invading AI technologies on citizens can entrench bias, inequality and undermine human rights, such as the right to privacy.

5.1 A multi-stakeholder perspective on trust in AI

Our concept matrix outlines the varying vulnerabilities of key stakeholder groups in relation to AI systems. Accepting vulnerability is a key element of trust and understanding and mitigating the risks and vulnerabilities AI systems pose for stakeholders, is central to facilitating trust and building the confident positive expectations that it is founded on. Given this we propose future research take a multi-stakeholder approach to examining the antecedents of trust in AI.

Prior research has shown that stakeholders' varying vulnerabilities in trusting another entity influence the salience and importance of the cues and antecedents that inform trust [35]. Understanding the vulnerabilities and expectations of different stakeholders of complex socio-technical systems is also important [100] because stakeholder alignment facilitates trust in firms seeking to innovate with AI [101].

However, as shown in our review, much of the research to date has focused on a single stakeholder, usually an individual end user or domain expert. A reason for this may be that most empirical research on the antecedents of trust in AI is experimental, and places participants either as quasi-users or a non-specific stakeholder role. Further, trusting behavior, and the antecedents that influence it, may be different in an experimental setting than in the field due to the varying risks, vulnerabilities and trust cues. For example, it is likely people will behave differently in an autonomous vehicle on the road than in a 'safe' driving simulator.

Moving forward, we see field experiments and longitudinal case studies examining multiple stakeholders of an AI system, as fruitful methodological approaches to deepen understanding of the antecedents of stakeholder trust in AI systems. Undertaking longitudinal case studies has the advantage of providing holistic, contextualised insights into the development of trust in AI systems over time. This is likely to provide a more systemic understanding of hitherto underexplored areas such as how stakeholder groups converge and diverge in relation to their vulnerabilities, expectations and trust in AI.

It is evident from our review that although several trust challenges have been raised, many have not been examined empirically, and few have been examined from the perspective of multiple stakeholders, or the perspective of society as a stakeholder.

Furthermore empirical studies have tended to examine whether a concept (such as accuracy or anthropomorphism) enhances trust, yet high trust is not always appropriate, and encouraging people to trust without 'good reasons' [33] can be manipulative. This tension is particularly apparent in studies of explainability and transparency, and anthropomorphism. For instance, people can misplace trust in inaccurate AI systems when provided an explanation [46], even nonsensical explanations [47], and anthropomorphism can lead people to believe that an agent is competent, even in the face of limited 'good reasons' [73]. Broadly, these issues can lead to overtrust and consequent problems. Further research is required to understand what influences stakeholders to trust 'optimally', that is in a well calibrated manner that aligns with actual evidence of trustworthiness and effective AI design that mitigates and minimizes the likelihood of harmful consequences [102].

6. References*

[1] K. Grace et al., "When will AI exceed human performance? Evidence from AI experts", Journal of Artificial Intelligence Research, 62, 2018, pp.729-754.

[2] AI HLEG, "Ethics Guidelines for Trustworthy AI", European Commission, 2018. Retrieved from https://ec.europa.eu/

[3] M. Söllner et al., "Trust", in MIS Quarterly Research Curations, Ashely Bush and Arun Rai, Eds., http://misq.org/research-curations, October 31, 2016.

[4] S. Makridakis. "The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms", Futures 90, 2017, pp. 46-60.

[5] Algorithm Watch, "AI Ethics Guidelines Global Inventory", 2018. Retrieved from https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/

[6] S.Y. Komiak and I. Benbasat, "The effects of personalization and familiarity on trust and adoption of recommendation agents", MIS Quarterly, 2006, pp. 941-960.

[7] L. Qiu and I. Benbasat, "Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems", Journal of Management Information Systems, 25(4), 2009, pp. 145-182.

[8] E.M Payne et al., "Mobile banking and AI-enabled mobile banking", Journal of Research in Interactive Marketing, 12(3), 2018, pp. 328-346.

[9] I. Panagiotopoulos and G. Dimitrakopoulos, “An empirical investigation on consumers’ intentions towards autonomous driving”, Transportation Research Part C: Emerging Technologies, 95, 2018, pp. 773-784.

[10] T. Zhang et al.,“The roles of initial trust and perceived risk in public’s acceptance of automated vehicles”, Transportation Research Part C: Emerging Technologies, 98, 2019, pp. 207-220.

[11] US Chamber of Commerce, “US Chamber of Commerce Principles on Artificial Intelligence”, 2019. Retrieved from https://www.uschamber.com

[12] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, “Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems”, 2019. Retrieved from https://standards.ieee.org/

[13] M.T. Ribeiro et al., “Why should I trust you?: Explaining the predictions of any classifier”, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, September 2016, pp. 1135-1144.

[14] K. Siau and W. Wang, “Building trust in artificial intelligence, machine learning, and robotics”, Cutter Business Technology Journal, 31(2), 2018, pp. 47-53.

[15] E. Glikson and A.W. Woolley, “Human trust in Artificial Intelligence: Review of empirical research”, Academy of Management Annals, advanced online publication, 2020.

[16] T. Zhang et al., “Service robot feature design effects on user perceptions and emotional responses”, Intelligent service robotics, 3(2), 2010, pp.73-88.

[17] T.W. Bickmore et al.,, “Tinker: a relational agent museum guide”, Autonomous agents and multi-agent systems, 27(2), 2013, pp. 254-276.

[18] K.S. Haring et al., “The influence of robot appearance and interactive ability in HRI: a cross-cultural study”, In International conference on social robotics, 2016, pp. 392-401. Springer, Cham.

[19] P.A. Hancock et al., “A meta-analysis of factors affecting trust in human-robot interaction”, Human Factors, 53(5), 2011, pp.517-527.

[20] K.E. Schaefer et al, “A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems”. Human Factors, 58(3):, 2016, pp. 377-400.

[21] OECD, “Artificial Intelligence on Society”, OECD Publishing, Paris, 2019. Retrieved from https://www.oecd-ilibrary.org/.ors, 58(3), 2016, 377-400.

[22] P.M. Krafft et al., “Defining AI in Policy versus Practice”, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 72-78.

[23] M.I. Jordan and T.M. Mitchell, “Machine learning: Trends, perspectives, and prospects”. Science, 349(6245), 2015, pp. 255-260.

[24] E. Brynjolfsson and A. Mcafee “The business of artificial intelligence”. *Harvard Business Review*, July 2017, pp. 1-20.

[25] Y. LeCun et al., “Deep learning”. Nature, 521(7553), 2015, p. 436-444.

[26] I. Rahwan et al., “Machine behaviour”, Nature, 568(7753), 2019, pp. 477-486.

[27] Y.K. Dwivedi et al., “Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy”, International Journal of Information Management, 101994, 2019, pp.

[28] P. Ping Li, “When trust matters the most: The imperatives for contextualizing trust research”. Journal of Trust Research 2(2), 2012, pp. 101-106.

[29] M. Deutsch, “The effect of motivational orientation upon trust and suspicion”. *Human Relations,13*(2), 1960, pp. 123-139.

[30] D. Gefen et al., “Trust and TAM in online shopping: An integrated model”, MIS Quarterly, 27(1), 2003, pp. 51-90.

[31] D.M. Rousseau, et al., “Not so different after all: A cross-discipline view of trust”, Academy of Management Review, 23(3), 1998, pp. 393-404.

[32] D.H. McKnight et al., “Trust in a specific technology: An investigation of its components and measures”, ACM Transactions on management information systems (TMIS), 2(2), 2011, pp. 1-25.

[33] J.D Lewis and A. Weigert, “Trust as a social reality”. Social Forces, 63(4), 1985, pp. 967-985.

[34] C.A. Fulmer and M.J. Gelfand, “At what level (and in whom) we trust: Trust across multiple organizational levels”, Journal of Management, 38(4), 2012, pp. 1167-1230.

[35] M. Pirson and D. Malhotra, “Foundations of organizational trust: What matters to different stakeholders?”, Organization Science., 22(4), 2011, pp.1087-1104.

[36] R.C. Mayer et al., “An integrative model of organizational trust”, Academy of Management Review, 20(3), 1995. pp. 709-734.

[37] D.E. Leidner, “Review and theory symbiosis: An introspective retrospective”, Journal of the Association for Information Systems. 19(6), 2018, Article 1.

[38] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)”, IEEE Access, 6, 2018, pp. 52138-52160.

[39] A. Holzinger, et al., “What do we need to build explainable AI systems for the medical domain?”, arXiv preprint arXiv1712.09923, 2017.

[40] R. Elshawi et al., “On the interpretability of machine learning-based model for predicting hypertension”, BMC medical informatics and decision making, 19(1), 2019, pp. 1-32.

[41] C. Gretton, “Trust and Transparency in Machine Learning-Based Clinical Decision Support”. In J. Zhou and F. Chen (eds.), Human and Machine Learning, 2018, pp. 279-292. Springer, Cham.

[42] P. Andras et al., (2018) “Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems”. IEEE Technology and Society Magazine, 37(4), 2018, pp. 76-83.

[43] V. Hollis, et al., “On being told how we feel: how algorithmic sensor feedback influences emotion perception”. Proceedings of the ACM on Interactive,

Mobile, Wearable and Ubiquitous Technologies, 2(3), 2018, pp. 1-31.

[44] A. R. Wagner et al., "Overtrust in the robotic age". Communications of the ACM, 61(9), 2018, pp. 22-24.

[45] J. A. Kroll, "The fallacy of inscrutability". Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 2018, 20180084.

[46] A. Weller, "Challenges for transparency", arXiv preprint arXiv,1708.01870, 2017.

[47] E.J. Langer et al., "The mindlessness of ostensibly thoughtful action: The role of" placebic" information in interpersonal interaction", Journal of Personality and Social Psychology, 36(6), 1978, pp. 635-642.

[48] J. Y. Chen, "Human-autonomy teaming and agent transparency", Companion Publication of the 21st International Conference on Intelligent User Interfaces, March 2016, pp. 28-31.

[49] C. S. Calhoun, et al., (2019). "Linking precursors of interpersonal trust to human-automation trust: An expanded typology and exploratory experiment", Journal of Trust Research, 9(1), 2019, pp. 28-46.

[50] E. S. Vorm, "Assessing Demand for Transparency in Intelligent Systems Using Machine Learning", 2018 Innovations in Intelligent Systems and Applications, July 2018, pp. 1-7.

[51] G. Du and G. Ruhe, "Does explanation improve the acceptance of decision support for product release planning?", In 2009 3rd International Symposium on Empirical Software Engineering and Measurement, October 2009, pp. 56-68. IEEE.

[52] R. F. Kizilcec, "How much information? Effects of transparency on trust in an algorithmic interface", Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, May 2016, pp. 2390-2395.

[53] H. Wallach, "Computational social science≠ computer science+ social data", Communications of the ACM, 61(3), 2018, pp. 42-44.

[54] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification". Conference on Fairness, Accountability and Transparency, January 2018, pp. 77-91.

[55] M. Yin, et al., "Understanding the effect of accuracy on trust in machine learning models", In Proceedings of the 2019 chi conference on human factors in computing systems May 2019, pp. 1-12).

[56] E. J. de Visser et al.,. "Learning from the slips of others: Neural correlates of trust in automated agents". Frontiers in Human Neuroscience, 12(309), 2018, pp. 1-15.

[57] A. Freedy, et al., "Measurement of trust in human-robot collaboration". 2007 International Symposium on Collaborative Technologies and Systems, May 2007, pp. 106-114.

[58] S. Moran et al., "Team reactions to voiced agent instructions in a pervasive game", In Proceedings of the 2013 international conference on Intelligent user interfaces, May 2013, pp. 371-382.

[59] M. Salem et al., "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust", 10th ACM/IEEE International Conference on Human-Robot Interaction, March 2015, pp. 1-8.

[60] S. Raisch and S. Krakowski, "Artificial Intelligence and Management: The Automation-Augmentation Paradox", Academy of Management Review. 2020, in press.

[61] V. Diebolt et al., "Artificial intelligence: Which services, which applications, which results and which development today in clinical research? Which impact on the quality of care? Which recommendations?" *Therapie*, *74*(1), 2019, pp. 155-164.

[62] K.T. Chui et al., "Big data and IoT solution for patient behaviour monitoring", Behaviour & Information Technology, 38(9), 2019, pp. 940-949.

[63] A.C. Valdez et al., "Recommender systems for health informatics: state-of-the-art and future perspectives", In Machine Learning for Health Informatics, 2016, pp. 391-414. Springer, Cham.

[64] C. Longoni et al., "Resistance to medical artificial intelligence". Journal of Consumer Research, 46(4), 2019, pp. 629-650.

[65] A. Lui A and G.W. Lamb, "Artificial intelligence and augmented intelligence collaboration: regaining trust and confidence in the financial sector", Information & Communications Technology Law., 27(3), 2018, pp. 267-283.

[66] E. de Visser E and R. Parasuraman, "Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload", Journal of Cognitive Engineering and Decision Making, 5(2), 2011, p.209-231.

[67] T. Rinta-Kahila et al., "Consequences of Discontinuing Knowledge Work Automation-Surfacing of Deskilling Effects and Methods of Recovery", In Proceedings of the 51st Hawaii International Conference on System Sciences, January 2018, pp. 5244-5253.

[68] S.G. Sutton et al., "How much automation is too much? Keeping the human relevant in knowledge work", Journal of Emerging Technologies in Accounting, 15(2): 2018, pp. 15-25.

[69] A.K. Heid, "Trust in Homo Artificialis: Evidence from Professional Investors", 2018 American Accounting Association Annual Meeting, August 2018, Washington USA.

[70] X. Fan X et al., "The influence of agent reliability on trust in human-agent collaboration", In Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction, January 2008, pp. 1-8.

[71] J. M. Logg et al., "Algorithm appreciation: People prefer algorithmic to human judgment". Organizational Behavior and Human Decision Processes, 151, 2019, 90-103.

[72] A. Waytz, et al., "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle". Journal of Experimental Social Psychology, 52, 2014, pp. 113-117.

[73] K. E Culley and P. Madhavan, “A note of caution regarding anthropomorphism in HCI agents”. Computers in Human Behavior, 29(3), 2013, pp. 577-579.

[74] A. L. Baker et al., “Toward an Understanding of Trust Repair in Human-Robot Interaction: Current Research and Future Directions”. ACM Transactions on Interactive Intelligent Systems, 8(4), 2018, Artilce 30.

[75] A. Salles et al., “Anthropomorphism in AI”, AJOB Neuroscience, 11(2), 2020, pp. 88-95.

[76] F.M. Verberne et al., “Trusting a virtual driver that looks, acts, and thinks like you”, Human factors, 57(5), 2015, pp. 895-909.

[77] K. Kim et al., “Does a digital assistant need a body? The influence of visual embodiment and social behavior on the perception of intelligent virtual agents in AR”, In 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), October 2018, pp 105-114. IEEE.

[78] W.A. Bainbridge et al., “The effect of presence on human-robot interaction’, InRO-MAN 2008 – The 17th IEEE Internationhal Symposium on Robot and Human Interactive Communication, August 2008, pp. 701-706. IEEE.

[79] B. Huynh et al., “A study of situated product recommendations in augmented reality”, In 2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), December 2018, pp. 35-43. IEEE

[80] J. Złotowski et al., “Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy”, Paladyn, Journal of Behavioral Robotics, 7, 2016, 55-66.

[81] S. Gupta et al., “Big data with cognitive computing: A review for the future”, International Journal of Information Management, 42, 2018, pp. 78-89.

[82] S. Zuboff, “Big other: surveillance capitalism and the prospects of an information civilization”, Journal of Information Technology, 30(1), 2015, pp. 75-89.

[83] E. Osnos, “Can Mark Zuckerberg fix Facebook before it breaks democracy?”, The New Yorker, 10 September 2018, Retrieved from https://www.newyorker.com/magazine/2018/09/17/can-mark-zuckerberg-fix-facebook-before-it-breaks-democracy.

[84] S. Kokolakis, “Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. Computers & Security, 64, 2017, 122-134.

[85] J. Andrew & M. Baker, “The General Data Protection Regulation in the Age of Surveillance Capitalism”, Journal of Business Ethics, in press.

[86] J. Vallverdú and D. Casacuberta, “Ethical and technical aspects of emotions to create empathy in medical machines”, In Machine Medical Ethics, 2015, pp. 341-362. Springer, Cham.

[87] I.R. Kerr and M. Bornfreund, “Buddy bots: How turing's fast friends are undermining consumer privacy”, Presence: Teleoperators & Virtual Environments, 14(6), 2005, pp. 647-655.

[88] T. Braun, B.C. Fung, F. Iqbal and B. Shah, “Security and privacy challenges in smart cities”, Sustainable cities and society, 39, 2018, pp. 499-507.

[89] K. Kaur and G. Rampersad, “Trust in driverless cars: Investigating key factors influencing the adoption of driverless cars”. Journal of Engineering and Technology Management, 48, 2018, pp. 87-96.

[90] S. Reig et al., “A field study of pedestrians and autonomous vehicles”, In Proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications. September 2018, pp. 198-209.

[91] N.L. Tenhundfeld et al.,”Calibrating trust in automation through familiarity with the autoparking feature of a Tesla Model X”, Journal of Cognitive Engineering and Decision Making, 13(4), 2019, pp. 279-294.

[92] C.W. Ho et al., “Governance of automated image analysis and artificial intelligence analytics in healthcare”, Clinical Radiology, 74(5), pp. 329-337.

[93] A.F. Winfield and M. Jirotka, “Ethical governance is essential to building trust in robotics and artificial intelligence systems”, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 2018, 20180085.

[94] S. Lockey et al., “Trust in Artificial Intelligence: Australian Insights”, The University of Queensland and KPMG Australia, October 2020.

[95] M. Guihot et al., “Nudging robots: Innovative solutions to regulate artificial intelligence”, Vand. J. Ent. & Tech. L., 20, 2017, p3p. 385-456.

[96] B. Zhang and A. Dafoe, “Artificial intelligence: American attitudes and trends”. Available at SSRN 3312874. 2019.

[97] Centre for Data Ethics and Innovation, “AI and Personal Insurance”, CDEI Snapshot Series, September 2019.

[98] X. Xhang, “Information Asymmetry and Uncertainty in Artificial Intelligence”, Medium, 9 September 2017, Retrieved from https://medium.com/@zhxh/information-asymmetry-and-uncertainty-in-artificial-intelligence-ad8e444c4d9a

[99] P. Nemitz, “Constitutional democracy and technology in the age of artificial intelligence”, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 2018, 20180089.

[100] S.A. Wright and A.E. Schultz, “The rising tide of artificial intelligence and business automation: Developing an ethical framework”, Business Horizons, 61(6), 2018, pp. 823-832.

[101] M. Hengstler et al., “Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices”. Technological Forecasting and Social Change, 105, 2016, pp. 105-120.

[102] E.J. de Visser et al., “Towards a Theory of Longitudinal Trust Calibration in Human-Robot Teams”, International Journal of Social Robotics, 12, 2020, pp. 459-478.

*Due to space constraints, we present selected references from our review. Contact SL for a full reference list.