Principles for understanding trust in artificial intelligence

Jim A.C. Everett[1†], Scott Claessens[1], Tim-Dorian Knöchel[1], and Madeline G. Reinecke[2]

[1] School of Psychology, University of Kent, Canterbury, Kent, United Kingdom
[2] Uehiro Oxford Institute, University of Oxford, Oxford, United Kingdom

[†]email: J.A.C.Everett@kent.ac.uk

Artificial intelligence (AI) increasingly performs tasks once reserved for humans, raising questions about when, why, and how people trust machines—and whether they should in the first place. In this Review, we identify six principles that help structure understanding of trust in AI and highlight its socially embedded nature: trust in AI is inferred; trustworthiness, trust, and trusting behaviour are distinct; trust in AI is about both morality and performance; and that trust in AI is agent-specific; individually variable; and strategically motivated. The inferred, multidimensional, dynamic, and contextual nature of trust in AI illustrates that 'trust in AI' is not one thing, but varies across different systems, individuals, and contexts. We end by considering broader ethical implications of studying trust in AI and argue that trust in AI requires both studying how people think and reflecting on the kind of world that trust in AI serves to create.

Citation

Everett, J.A.C., Claessens, S., Knöchel, T.D., & Reinecke, M. (2026). Principles for understanding trust in artificial intelligence. *Nature Reviews Psychology.*

Acknowledgements

This work was generously supported by funding from a UKRI Horizon Guarantee (EP/Y00440X/1) awarded to JACE. MGR is supported in part by the Wellcome Trust [226801]

Author contributions

JACE: conceptualisation, writing original draft, writing final draft, funding acquisition. SC, TK, and MGR: writing original draft, writing final draft.

Principles for understanding trust in artificial intelligence

Artificial intelligence (AI) systems are increasingly being used to assist—or even take over—tasks that were once the preserve of humans. It is therefore no surprise that questions about trust in AI have moved to the centre of public and scientific debate. Computer scientists are building AI in ways they hope will mean it is trusted to ensure adoption and drive uptake[1]. Psychologists are studying the antecedents of trust in AI[2,3], validating self-report scales of trust in AI [4,5], and exploring how people trust and use AI in applied settings[6,7]. Ethicists are outlining principles for 'trustworthy AI' to ensure that AI is safe and effective[8,9] and the European Union has official guidelines for building trustworthy AI[10].

Determining when, why, and how people trust machines, what trust in AI means in the first place and what kind of trust is desirable, however, is far from straightforward, in part because 'trust' can mean different things (Box 1). 'Trust' includes an assessment of performance or reliability, but also more psychologically-driven inferences about intentions and good motivations. 'Trustworthiness' is both a normative assessment of whether trust is warranted and a psychologically inferred process that leads to an attitude and behaviour of trust. People do not trust all AI systems equally, not all people trust the same AI system in the same way, and even the same person might trust the same AI system differently in different contexts based on their strategic motivations and goals. All of these factors play out against a broader ethical background concerning—perhaps paradoxically—whether trust in AI is the right vision for society.

In this Review, we combine classic perspectives on trust with contemporary perspectives and empirical findings emerging in a rapidly evolving landscape of AI trust research. Trust in AI has been widely considered across multiple disciplines including psychology, philosophy, and computer science. Large-scale reviews and meta-analyses have already compiled literature on trust in AI generally[2,11–15] as well as in aversion to AI[16,17], AI in moral decision-making[18,19], AI in managerial decision-making[20,21], and trust in autonomous vehicles[22]. Here we take a different approach. Instead of attempting the Sisyphean task of presenting all available literature in a fast-moving landscape, we identify guiding principles to help structure insight into the topic and understand the way it is changing. First, we discuss the definition of 'trust in AI'. Next, we combine insights from psychology and other disciplines to identify six principles that can guide understanding of trust in AI. These principles are not intended as hard-and-fast rules, but as a framework for considering trust in AI and avenues for future research. Fundamentally, trust in AI is not a single attitude or feature, but a dynamic, context-dependent, and socially embedded process. We then consider the ethical implications of studying trust in AI and conclude with directions for future research.

Defining 'trust in AI'

'Artificial intelligence' can mean different things to different people. Here we adopt a typical understanding of AI as referring to machines or systems that use any kind of algorithm or statistical model to perform tasks or make decisions that usually require human intelligence[23].

'Trust in AI' can also mean different things. For example, when someone says they trust a specific AI tool, they might be expressing a belief that the AI has good performance and does things reliably and competently; they could mean that they think they could rely on the AI in cases of uncertainty; they could mean something richer, like when people say they trust another person, meaning they think the AI is sincere or 'good' in some way; or they could mean that they actually use the tool and it has some influence on their behaviour.

In the psychology literature, trust is most commonly understood as "the willingness of a party to be vulnerable to the actions of another party"[24], or "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action"[25]. Drawing on these ideas, one of the most-commonly cited definitions of 'trust in AI' defines trust as "the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability"[26].

Although researchers and the public discuss trust in AI, some philosophers have argued that trust in AI is "conceptual nonsense"[27], because "people trust people, not technology"[28]. The standard philosophical claim (which comes out of decades of philosophical work on trust[29,30]) is that genuine trust requires rich interpersonal attitudes between a trustor and a trustee. These depend on agentic capacities, such as being motivated by good will or responding to normative commitments[31–33]. Thus, although AI can be relied upon just as someone might be able to rely on a chair or a computer, it does not make sense to talk about genuinely trusting AI: AI simply does not have the kind of mental states and motivations required for the interpersonal relationship of trust[31,32]. According to such arguments, trust in AI is merely reliance on AI, and the idea of trust in AI is mistaken, misleading, or even potentially dangerous[27,32–34].

Even if these philosophical arguments are correct and AI is not the kind of thing that could genuinely participate in an interpersonal relationship of trust, it is an entirely separate question whether people nevertheless treat AI systems as social agents that can be meaningfully trusted. Indeed, evidence suggests that they do. In line with the computers as social actors framework[35–37], people apply social rules and expectations to machines[38] and perceive computer personalities[39]. People also anthropomorphise AI in much the same way that they anthropomorphise other non-human agents[40], and conform their moral judgements with AI avatars in virtual reality, just like they do when making moral judgements with a group of other humans[41]. Most pertinently, a preprint suggests that people do think AI is the kind of thing that

can have good intentions and be trusted, like other social agents, with participants reporting that it makes sense and is natural to say "I trust the AI" but not "I trust the moon" or "I trust the number five"[42].

In sum, 'trust in AI' can mean different things, and is perhaps even, strictly speaking, a conceptual mistake[32]. But it is also psychologically meaningful, intensely debated, and therefore pragmatically indispensable for understanding the ways in which people interact with AI systems.

Principles for understanding trust in AI

In this section we identify six principles that can guide understanding of trust in AI. First, we discuss how trust in AI cannot be engineered into a machine because it is inferred. Second, we highlight how trustworthiness, trust, and trusting behaviour are connected but distinct constructs. Third, we show that trust in AI is multidimensional and includes assumptions of competence and ability as well as broader perceptions of the system's morality. Fourth, we show the richness of understanding trust in AI is likely to come not from generalities, but from recognising the agent-specificity of trust in AI: just as people do not trust all humans in the same way, people do not trust all AI systems in the same way. Fifth, we discuss how even the same AI system can be trusted in different ways by different people because trust is influenced by various individual and cultural differences. Finally, and most underexplored in the current literature, we discuss how people might trust AI systems differently depending on their specific strategic goals and motivations in different kinds of interactions.

Trust in AI is inferred

Trustworthiness can be understood as a normative, evaluative question: Is this entity worthy of trust? In the context of AI, the normative aspect of trustworthiness often leads to a prevailing assumption in AI ethics and governance that trustworthiness is something that can and should be engineered into machines to enhance trust. The European Commission's *Ethics Guidelines for Trustworthy AI* has argued that "trustworthiness is a prerequisite for people and societies to develop, deploy and use AI systems"[10]. Scholars have highlighted numerous principles for what would make AI trustworthy[43], including the European Commission guidelines positing that a trustworthy system is one that is lawful, ethical, and robust. This framing assumes that trustworthiness is something that can be built. Consequently, computer scientists, developers, and regulators have become interested in understanding how these principles can be translated into practice[1,44].

Some work treats trustworthiness as an objective characteristic. For example, in AI ethics, trustworthiness is conceptualised as a normative question that refers to capacities of the system, such that trustworthiness is "an objective attribute of the trustee", distinct from trust which is "a subjective attitude and attribute of

the vulnerable party"[45]. By contrast, other work treats trustworthiness as a psychological and inferred characteristic. For example, much of the psychological literature on trust in AI draws on an influential model of organisational trust[25] and treats trustworthiness as a perceived assessment of the characteristics of the agent that forms the basis of trust[3,26,46].

Importantly, whether a system objectively has characteristics that make it worthy of trust (actual trustworthiness) and whether people perceive that the system has those characteristics (perceived trustworthiness) are separate. A system can be actually trustworthy while not perceived as trustworthy, and a system can lack actual trustworthiness but still be perceived as trustworthy[47,48]. Thus, although engineers might be able to program AI systems that are objectively safe and reliable, they cannot program trust. Indeed, trust is not something that can be identified in computer code or governance regulations because it is a product of human impression formation. These impressions are influenced by objective features of the system, but these objective features are interpreted and inferred by humans.

Psychology research shows that people form judgements of competence and morality from minimal cues to build an overall picture of someone's character[49,50]. These inferences are made heuristically and often automatically[51,52] and people can be misled by irrelevant features like attractiveness[53]. When AI systems are posited to be and treated as social or quasi-social actors, people extend these same inferential mechanisms to them[54]. Thus, people form implicit and explicit impressions of AI[55] based on limited information that is not necessarily driven by an AI's actual trustworthiness. Even the assessment of seemingly objective AI system characteristics, such as its accuracy, varies between observers[56].

One example of a mismatch between actual trustworthiness and perceived trustworthiness comes from work on interpretability. Many contemporary AI systems can work as black boxes with internal processes that are largely inscrutable. This lack of transparency has raised concerns about the ethical risks and safety of using such systems, which has led to a focus on designing explainable AI[57]. But interpretability will not necessarily increase trust in practice because trust is an inferred process, not a feature of the machine. Indeed, evidence on whether interpretability actually increases trust in AI is mixed[58–61]. These results highlight the tension between what should make a system trustworthy and what people actually trust.

In sum, developers can focus on building normatively trustworthy systems, which is likely to increase people's trust in their products. However, these points are distinct: the fact that something has characteristics that increase its actual trustworthiness does not guarantee people will perceive it as such, and the fact that people perceive a system as trustworthy does not mean it is actually trustworthy.

Trustworthiness, trust, and trusting behaviour are distinct

Most psychology research on trust in AI draws on the tripartite model of interpersonal trust[25], which has since been adapted to the AI context[62]. This model can be drawn on to distinguish between perceived trustworthiness, trust, and trusting behaviour: an assessment of trustworthiness leads to an attitude of trust, which can in turn translate to a behaviour of trust (Figure 1). Correspondingly, trustworthiness is typically understood as the basis of trust, or as a perceived assessment of the characteristics of the agent that forms the basis of trust[3,26,46], whereas trust is an attitude or belief indicating a willingness to rely on the agent[25]. Some research subdivides this attitude or belief further, and distinguishes rational trust that is based on evidence of the other person's capacities (cognitive trust) from a more irrational expression of trust that is based on an emotional response to the other person (affective trust)[63].

However, the distinction between trustworthiness and attitudinal trust is complicated by how the latter is typically measured. Although attitudinal trust can be measured by direct self-report with questions like "Do you trust ChatGPT?", it is also measured by a plethora of multi-item self-report scales (Table 1) that often ask participants both about whether they trust AI and about the perceived capabilities or trustworthiness of the system (for example, "Is ChatGPT reliable?")[4,5]. Such questions are clearly about the characteristics that would form a basis of trust (perceived trustworthiness), but they also reflect an overall attitude that can be referred to as trust itself.

Indeed, it is difficult—and perhaps not even desirable—to distinguish expectations of the capacities of a system from trust simpliciter[4]. Oftentimes attitudinal trust is not separate from an assessment of performance, and the performance assessment itself can be meaningfully understood as trust[4]. At the same time, however, it is clear that attitudinal trust is not identical to perceived trustworthiness because even perceptions of an AI's ability do not perfectly correlate with ratings of attitudinal trust and other factors, such as general propensity to trust[64]. 'Trust' can therefore refer to both an assessment of the capacities of the AI system as well as a self-reported willingness to rely on the system.

Trust in AI can also manifest behaviourally[65]. Trusting behaviour is captured with behavioural measures such as people's decisions to rely on an AI system[66], delegate to a system[67], or update their judgements following a system's recommendations[68]. The latter is often discussed as persuasion, not a manifestation of behavioural trust per se[69,70], highlighting a potential jingle-jangle fallacy[71–73] (using the same term to mean different things and/or different terms to mean similar things). Some researchers have measured behavioural trust in AI by adapting the incentivised trust game[74] from experimental economics in which one player (trustor) sends some amount of money to another player (trustee), who then decides how much money to return to the first player. In the adapted game, an AI is placed in the position of the trustee instead of a human[75–77]. Although the trust game has been widely used as a behavioural measure of trust (how much the trustor gives to the trustee) and trustworthiness (how much the trustee returns) between humans, AI systems are unlikely to value money in the same way that humans do, which highlights

potential challenges with applying standard behavioural methods for assessing human-human interactions to human-AI interactions.

According to the classic model of interpersonal trust, perceived trustworthiness leads to attitudinal trust, which then leads to trusting behaviour[25] (with actual trustworthiness feeding into perceived trustworthiness). But this is far from a rule. Risk and stakes play a role in moving from attitudes to behaviour[25], and researchers have long-known that behaviour does not always directly follow from attitudes[78,79]. Indeed, much research has highlighted situations where people report that they trust an AI system but do not actually use the technology in practice[48,65,80]. This disconnect between trusting attitudes and behaviour might occur when people are anxious about new technologies or perceive high levels of risk in using the AI system[81–83]. There is also emerging evidence for an opposite and less explored pattern: that people might hold explicit attitudes about AI being untrustworthy, but still engage in trusting behaviour such as updating their judgements based on AI moral advice[84]. A cross-cultural study presented in a preprint found that people reported trusting AI moral advisors less than human advisors, but updated their judgements similarly based on advice from AI and humans[68]. These findings underscore the need for conceptual clarity in the measurement and discussion of trust in AI, as well as consideration of how increasing use of AI could reshape classic debates about attitude-behaviour mismatches.

In sum, 'trust' can refer to an assessment of the capacities of an agent, an expression of a willingness to rely on it in cases of uncertainty, and a behavioural manifestation. All these meanings are justifiably referred to under the umbrella of 'trust' and are included in measures of trust, but they can also meaningfully diverge (Figure 2). The difficulties associated with broad and competing understandings of trust and trustworthiness is a barrier in the literature on trust in AI[48,65,80]. It is therefore important that researchers are clear about what is meant by 'trust' (assessments of trustworthiness or capacities, a self-reported attitude about trust, or trusting behaviour).

Trust in AI is about both morality and performance

When a user asks themselves whether they should trust an AI system, in one sense they are simply asking "Can I rely on this system to function reliably and accurately?". This dimension of trust that focuses on reliability, ability, or performance (performance trust) has been suggested to be the primary driver of trust in AI[15,26,85]. Indeed, the perceived performance capability of an AI system is a key predictor of trust in that system[85], and trust in an AI system quickly declines when algorithms are shown to perform incorrectly, especially on tasks that are considered easy[86,87].

It might have made sense to focus on performance trust for early AI systems that had limited functionality. However, the rise of more complex and socially-situated AI systems has made clear that thinking about

trust in AI requires considering more than performance capabilities. For example, with enhanced sensors autonomous weapons could be more efficient than humans on the battlefield. However, people strongly distrust and disapprove of autonomous weapons being used in military strikes, largely because they are seen as less moral[88].

According to classic tripartite models of human interpersonal trust, trust and trustworthiness are based not only on perceived ability, but also on the integrity and benevolence of the trustee[46]. Others have identified competence and responsibility[89], ability and intentions[90,91], and expertise and motivation to lie[92], as dimensions of trust, in line with theories of impression formation showing that people care not only about whether a person is competent, but also whether they are warm, moral, and trustworthy[49,51].

Trust in AI is not solely dependent on perceptions of performance or ability ("Does it get the job done?"), but also concerns about moral trust[4], or the system's purpose and process[26] ("Is it 'good'?"). Research on trust in AI does not consistently distinguish between performance and ability vs. purpose and process (or morality)[15,93,94]. People can, do, and should care about the ability and performance of AI systems[4,26,85] because these characteristics enable the truster to have confidence in the trustee's skills. But people also can, do, and should care about the system's process and purpose[26], benevolence and integrity[46], and ethicality[4], that is, whether the perceived intentions and motivations of the trusted system are good. An AI system can be highly capable but deceptive, reliable but unethical, or ethically aligned yet unreliable.

One multidimensional model of trust in AI distinguishes between performance trust and moral trust[4]. As a psychological inference, moral trust does not require that the AI is a moral agent or genuinely has benevolence, integrity, or honesty. Instead, moral trust depends on whether people treat the AI as a moral agent and evaluate it not only for its performance but also for its potential to do harm. Thus, moral trust involves assessing whether the AI is ethical (follows moral norms and adheres to principles that seek to do good) and whether it is sincere (is genuine and authentic). Sincerity has received less attention than integrity and benevolence (here broadly analogous to the ethical subdimension) in work on interpersonal trust[4], and was potentially less relevant for early algorithmic systems. However, sincerity is increasingly important in discussions of generative large language model (LLM) products like ChatGPT given their reputations for 'lying' to users by producing factually false outputs[95].

Meta-analytic work shows that both performance and perceptions of benevolent rule-following influence trust in AI systems[2]. Other work shows that perceived ability, benevolence, and integrity of a virtual chatbot agent form a latent factor of trustworthiness that influences trust[3], that perceived authenticity of healthcare chatbots influences trust[96], and people perceive AI as more trustworthy both when its morality increases and when its performance increases[97].

Although there is moderate consensus that trust in AI can be driven by both performance and moral elements, the literature does not consistently distinguish between performance and moral dimensions of trust[15,93,94]. Similarly, many scales of trust in AI include items related to both morality and performance, broadly construed, although they are rarely referred to with the same terminology and are not always explicitly distinguished in separate dimensions or subscales (Table 1). For example, the widely-used trust in automation scale includes items related to both performance (for example, "The system is reliable") and morality (for example, "The system has integrity") within a single scale[5].

In sum, both morality and performance drive trust in AI, but they are also distinct: there is no guarantee that an AI that is more intelligent will necessarily be more moral[97–99], and no guarantee that perceiving an AI as more reliable means it is perceived as more ethical given the multidimensional nature of trust[4,25,100]. It will be important for future research to investigate the independent—and perhaps even competing—roles that perceptions of performance and morality play in shaping trust in AI.

Trust in AI is agent-specific

Much existing work has focused on people's trust in AI in general, and have identified principles of trustworthy AI that should be generally applicable to any autonomous system or agent[8,9]. However, there are many different types of AI systems with different characteristics that are used for different purposes, and it is far from clear that people's trust in general 'AI' develops and extends in the same way across, for instance, GPS navigation systems, LLM chatbots, and AIs used in drone warfare. Moreover, variables that are known to influence trust in AI in general (such as transparency or interpretability[57–60,101] and degree of anthropomorphism[102,103]) might have different effects on trust for different systems. For example, higher interpretability that enables a human to evaluate a machine's internal processes is likely to be more important for trust in an AI system that makes more consequential decisions about parole recommendations than an AI-system like DALL-E that creates images[58].

Given this variability in AI systems, the tendency to generalise across empirical findings has been argued to be one of the greatest challenges in the study of trust in AI[45]; the question 'Do people trust AI?' must be qualified: 'Trust which AI to do what?'. For example, a large global survey (n = 48,340) found that although almost half of respondents reported being willing to trust an AI overall, there were notable differences in their willingness to trust AI built for specific contexts[104], such that people trusted healthcare AI more than generative AI and AI used for human resource management. Other empirical studies show that people are more averse to machines making moral decisions than to machines making other kinds of decisions[105], and people even trust AI systems differently based on the AI's recommendations in moral dilemmas[68,106]. Thus, people's ratings of trust in AI generally can differ substantially from their rated trust in specific AI applications.

The agent-specificity of trust in AI means that people not only trust systems differently depending on their characteristics and domain of use, but that there are also different bases of trust for different systems. For example, a meta-analysis found that AI performance predicted trust across five categories of AI, but the performance of the system was relatively more predictive of trust in general algorithms and automated vehicles than trust in socially-situated robots and chatbots[2]. Similarly, according to a non-peer-reviewed preprint, although performance and moral indicators were positively related to overall trust across 20 AI systems, their relative importance varied such that performance was a stronger indicator of trust for some systems (such as Google Maps AI), whereas morality was a stronger indicator of trust for others (such as autonomous killer drones)[100].

Moreover, when AI systems are designed for different contexts, their perceived capabilities might or might not align with the traits that are expected for that specific context. For example, a survey of 10,000 participants across 20 countries found that people were more fearful of AI the more their expectations about its abilities were misaligned with the abilities required for the occupation in which it was being deployed[107]. Thus, trust in different AI systems is not only dependent on which capacities they have, but how these capacities match the context.

In sum, although much existing work has focused on trust in AI as a broad umbrella category, research shows that trust in AI is agent-specific. People trust specific AI systems differently depending on their particular functions, capacities, and domains of application.

Trust in AI is individually variable

People differ in their general propensity to trust others (for example, their agreement with the statement "most people can be counted on to do what they promise to do"), and this disposition tends to be higher in collectivist societies[108]. Such dispositional tendencies provide a foundation for how individuals form trust judgements towards humans and non-humans alike. However, individuals' propensity to trust automation (defined as a general tendency to trust technologies such as robots, autonomous vehicles, and AI-based systems) is correlated with, but conceptually distinct from, the propensity to trust other humans[109]. Specifically, it has been suggested that the propensity to trust other humans forms the broader dispositional basis from which trust in automated technology emerges[109]; this hierarchical relationship is supported by moderate correlations between the two constructs in England ($r \approx .45$; but see ref[110] for more modest correlations in Germany and Singapore).

Further large-scale evidence demonstrates the heterogeneity of individuals' trust in artificial systems. A systematic review of over 500 empirical studies across 62 countries identified a range of individual

characteristics (for example, personality traits, self-efficacy, technological experience and risk orientation) as among the most consistent antecedents of trust in AI[12]. In general, people high in extraversion, openness, and agreeableness tend to trust different AI systems more readily[111]. For example, people higher in extraversion report greater trust in AI-enabled user interfaces[112]; people higher in openness report greater trust in robots[113] and autonomous vehicles[114]; and people higher in agreeableness report greater trust in conversational AI[115]. It is therefore possible to identify personality profiles for potential users of autonomous vehicles[116]. In addition, people with higher AI competence self-report more trust in AI in general[117], whereas people with higher attachment anxiety report lower trust in different AI systems[118].

Other individual difference variables, such as age and gender, have also been examined as predictors of trust in artificial systems. However, meta-analytic evidence reveals inconsistencies or even null effects across studies[2,12]. This variation in the literature might reflect the fact that trust in AI is agent-specific. For instance, women trusted an autonomous 'robo-cop' more so than men[119], whereas a study of older adults' acceptance of 'assistive social robots' showed the opposite gender effect (more trust among men than women)[120].

Effects of age and gender also vary in size across cultures[121]. Indeed, there are robust cross-cultural differences in how people trust AI. For example, one systematic review found striking variation across countries' average trust in AI systems, with scores ranging from 47.39 in Spain to 79.80 in Malaysia (on a standardised scale from 0–100, with 100 indicating complete trust)[12]. These results reveal that global trust in AI is far from uniform. Participants in many Asian and Middle Eastern nations (including Malaysia, Singapore, and Iraq) reported higher baseline trust than participants in Western European and East Asian nations (such as Germany, Spain, and Japan). Participants in the U.S and Canada fell between these extremes, indicating cautious but generally optimistic attitudes toward AI. The authors attribute these differences to distinct cultural value systems and societal experiences with technology, such as different levels of uncertainty avoidance, power distance, and exposure to automation[12]. Other global investigations found higher trust in AI in countries with emerging economies[104] (such as Nigeria, India and Egypt) or lower prevalence of English (such as Indonesia, Malaysia and Thailand)[122]. These findings might reflect participants' trust in their governments to regulate AI development effectively.

Cross-cultural experiments further support the role of cultural values in trust in AI. For instance, although overall trust levels were similar between participants from the US and India, the underlying drivers of trust differed[123]: concerns about uniqueness neglect (the feeling that algorithms overlook individual distinctiveness) were associated with lower trust among US participants, whereas familiarity with the algorithm was associated with aversion and increased acceptance among Indian participants. These findings suggest that cultural frameworks can influence not only the extent to which people trust AI, but also why they come to do so.

In sum, trust in artificial intelligence can vary across a host of individual and cultural characteristics. All else being equal, a single AI system might inspire trust in some and skepticism in others. However, it is important to recognise that the results of cross-cultural surveys and experiments can be difficult to interpret because they might reflect sampling biases, differences in response styles, and a potential lack of measurement invariance for scale items across cultures.

Trust in AI is strategically motivated

It is often either implicitly or explicitly assumed that there is a linear relationship between trust and trustworthiness: an AI that is 'better' (or objectively worthy of trust) will be perceived as more trustworthy, which will generally lead to a higher attitude of trust. Such a linear relationship would be rational, but people take the risk of trust in a given situation into account[25] and, of course, humans rarely display full rationality. Instead, humans are influenced by a range of cognitive biases, and such biases even impact how people perceive risk[124]. In particular, people often exhibit motivated social cognition: they act less like scientists seeking the truth and more like lawyers confirming existing intuitions[125–127]. Applied to trust in AI, such motivated social cognition suggests that an individual might trust even the same AI differently depending on their strategic goals and motivations in a particular situation.

AI products like ChatGPT or 'AI mode' for Google Search might be intended to provide information, even if there are reasons to doubt the reliability of their answers[95]. How people seek out and use information can be influenced by strategic motivations, biases, or demands that are separate from the system itself. For instance, trust in AI is more likely when there are higher task demands that reduce cognitive resources[128,129]. This finding dovetails with an extensive literature on confirmation biases[130], according to which people pay more attention to information that favours their beliefs[131] and neglect information that goes against their initial view[132]. In the same way, people tend to search for information with AI in biased ways that can foster inaccurate beliefs[133]. A preprint suggests that these biases can be exacerbated by 'sycophantic' AI tools, which are also seen as more intelligent and moral[134]; other work finds that people are persuaded by AI giving moral advice in the absence of good reasons, as if they are applying a heuristic of 'this advice seems good enough'[84].

There are reasons to expect that people might seek out and trust AI in motivated ways in other cases beyond information retrieval. We know that using AI shapes perceptions of the user themselves, highlighting how trust in AI is not just about one what does, but the message it sends [135]. Although morality dominates impression formation in general[49], and people preferentially seek out information about others' moral character[136], there are exceptions: sometimes people might prefer to seek information about a partner's willingness to break rules or be less moral. For example, there is evidence that people

prioritise less moral partners if there is a strategic reason to do so, such as when selecting a defense lawyer[137]. Regarding trust in AI, evidence suggests people are more trusting of self-driving cars for other people than they are for themselves[138] and trust self-driving cars that protect themselves as passengers more than cars that protect pedestrians[139]. People are less likely to perceive an AI as immoral or unjust when they personally benefit from the AI's unfair decision[140], and some people might be more persuaded by an AI that gives selfish rather than altruistic advice[84]. Other work suggests that people are more likely to engage in dishonest behaviour in a die-rolling task when they can delegate their decision-making to an opaque algorithm[141]. These effects could be partly explained by the combination of psychological distance[142], plausible deniability[143], and deflection of moral responsibility[144] that AI systems offer their users. These findings raise concerns that people might not only trust AI because they perceive it as a good, reliable system but because that system aligns with their incentives, and might therefore over-rely on AI in situations where there is an incentive to engage in ethically questionable behaviour[18].

In sum, the idea that trust in AI systems might be strategically motivated has received less attention than the other principles. However, exploring how trust in AI is dependent on a motivated assessment of its alignment with strategic incentives, needs, and motives is a particularly promising area of future research. Even the same person could trust the same AI differently in different contexts based on their motives in that particular situation.

Ethical considerations

Trust in AI is a multifaceted, multidimensional, agent-specific, individually variable, and motivated phenomenon. By reviewing evidence from multiple disciplines, we have sought to document how, when, and why people might trust AI. But such evidence does not speak to how much people should trust AI in the first place, or how much researchers and policymakers should focus on increasing trust in AI.

AI clearly has the potential to do good in society. AI tools can be used to increase agricultural yields and improve food security[145–147], deliver faster disaster management in crises[148,149], monitor extreme weather and climate events[150], assist conservation efforts[151,152], and create breakthroughs in protein folding to discover new drugs[153,154]. Given these potential societal benefits, it is unsurprising that there is so much focus on trust in AI: trust is valued, important, and necessary in the human social world, and the potential benefits of AI tools will only be achieved if people trust those systems. However, trust is also dangerous when misplaced.

For all the potential benefits of AI systems, blind trust in AI is not the right goal, and it is important that researchers avoid a trap of assuming more trust is better[155]. Empirical work has documented concerning effects of over-trust in AI in different domains, including risky financial decisions[156], threat-identification

of enemy combatants vs. civilians[157], and food identification tasks[158]. It is therefore more important that AI is trustworthy (that is, the system has the capacities that make it worthy of trust) rather than trusted (that is, people perceive the system as good and are willing to rely on it)[34]. Just as people can be epistemically and ethically misled and even harmed by placing trust in other humans who are not worthy of that trust, people can be misled and harmed by placing trust in AI systems that are not worthy of that trust. Trust should therefore be appropriately calibrated such that an AI is trusted to the degree that the system is trustworthy[159] and designed to be relied upon in a particular situation[62]. For this reason, regulation and oversight are needed to ensure system limitations are acknowledged[160,161], and that this information is used to foster an appropriate level of trust[162].

However, even if researchers or policymakers focus on appropriate or calibrated trust there are deeper potential ethical implications of 'trust in AI' because the concept is inherently value-laden. Concerns have been raised about the commodification of trust[163], whereby the banner of 'trustworthy AI' can serve as a form of ethics-washing: corporations and governments can signal moral responsibility while avoiding substantive ethical or regulatory commitments or enquiry[27,164]. Critics have argued that frameworks such as the European Commission guidelines on trustworthy AI[10] function less as true safeguards and instead as "elegant public decorations for a large-scale investment strategy"[27]. Even badges or certification labels for 'trustworthy' AI could be misused by developers to drive unwarranted trust in a form of 'machine washing', just as companies can greenwash by presenting products as more environmentally friendly than they are[165].

Others have argued that describing AI as something that can be trusted assigns capacities to AI that it does not have, in essence implicitly treating AI as a moral subject with rights and responsibilities rather than a tool[42,166]. Framing AI as a trustworthy agent might obscure and shift responsibility away from AI developers to the systems themselves[167,168]. This form of 'agency laundering'[169] focuses attention on the responsibilities of AI as agents rather than the responsibilities of the developers, users, and regulators, and the way that AI systems can and do cause harm to real people[170]. Empirically, there is evidence that having AI recommendations led to more errors of commission and omission in simulated flight tasks than when AI recommendations were not provided, suggesting that people felt less accountable when responsibility can be assigned to AI than when it lies in the hands of the user alone[171]. Consistent with this interpretation, people perceived themselves to have less responsibility for decisions in a simulated drone strike task when they received AI assistance compared to when they did not[172].

When people learn about new concepts, labels assigned to that concept influence perceptions of concept valence[173], and research from cognitive linguistics suggests that when a familiar word is used in an unconventional manner, people tend to import features of the word's primary meaning[174,175]. Thus, the concept of 'trustworthy AI' might itself influence people's perceptions of trust and responsibility. A

preprint that has not been peer-reviewed suggests that people imbue AI with agential capacities that it does not have[42], suggesting that it is possible that considering AI as trustworthy agents shape perceptions of responsibility.

At the most abstract level, there are ethical debates about whether trust in AI—even appropriately calibrated trust—is desirable in the first place. Put differently, even if it became clear that a proprietary AI could be trusted to deliver teaching and grade student assignments to the same standard as human teachers, it is not automatically a given that the goal should be to ensure that students, parents, and administrations use these systems in lieu of human teachers. Scholars have raised concerns about the threat of 'algocracy', in which the proliferation of algorithm-based systems structure and constrain opportunities for human participation[176], lead to a technologically advanced but ethically deskilled society[177], and prioritise technological innovation at the expense of human flourishing[178].

Although rarely discussed in the context of psychological work, discussion of trust in AI can never be fully isolated from the context in which AI is developed and used—a context in which technology companies can use their money and influence to change policies to align with their own economic interests[179], where AI systems harm the most vulnerable people in society[180], where AI systems have stark environmental effects[181], and where the labour of training systems is often outsourced to underpaid and exploited workers in the global South[182]. These ethical and political contexts cannot be fully isolated from the task of understanding trust in AI because they constitute the conditions under which trust judgements are formed and expressed.

Psychologists studying trust are not—and should not be—held responsible for this ethically-fraught background. At the same time, it is impossible to fully divorce values from science, because ethical, political, and social assumptions shape not only which questions are asked, but how results are interpreted and used. For example, research that identifies features that increase trust in AI systems can easily be interpreted as guidance for designing systems that are more trusted, even though being more trusted does not correspond to being more trustworthy. Researchers studying trust in AI should therefore reflect on the potential uses and misuses of their findings, distinguish clearly between trust and trustworthiness in their framing of results, reflect on how their results exist within a broader context, and where suitable make explicit the assumptions and values that guide their research. Practices such as greater reflexivity about researchers' perspectives and, where appropriate, the use of positionality statements might help make these influences more transparent.

Summary and future directions

Drawing from research in psychology, philosophy, behavioural economics, human-computer interaction, and AI ethics, we formulated six key principles that underlie the burgeoning literature on trust in AI. The emerging picture is that trust in AI is not a static or unitary phenomenon. Instead, trust in AI is composed of multiple dimensions and constructs, is dynamically inferred, and varies substantially depending on the characteristics of the particular AI system, the individual doing the trusting, the cultural context, and the strategic affordances of the situation. Our Review is by no means exhaustive and the literature on trust in AI is growing every day. However, recognising the theoretical and empirical work underlying these principles and their interplay, as well as the ethical and practical challenges of studying trust in AI, has implications for researchers, developers, and policymakers (Table 2).

Although some aspects of trust in AI have been heavily studied, important areas are receiving less attention. There is a wealth of research on individual and cultural differences in trust in AI or characteristics that might influence trust in AI in general. However, there is much less research on people's strategic motivations for trusting AI, the way that performance and moral trust might diverge for different agents, or how characteristics of the AI will be differentially important for different systems based on the perceived degree of fit. Understanding of trust in AI could also be enhanced by exploring the interplay between the different principles we outlined here, such as how people might trust the same system in different ways, how even the same system may be trusted differently by different people in different contexts, and how the same predictor or characteristic may lead to trust differently across different systems. For example, one global report found that relative trust rankings for different AI systems vary substantially across cultures[104]. Future work should extend this finding by investigating whether there are individual and cultural differences in how people prioritise performance and moral trust in AI, the cues that people use to infer the trustworthiness of different AI systems, and people's strategic trust in AI.

Much of the research cited in this Review is cross-sectional. However, this is a fast-moving space, and new AI systems are released every day. The type of machine systems that discussions of automation focused on two decades ago[26], or that were included in studies of algorithm aversion just a decade ago[86], are very different from the kinds of LLMs that people routinely interact with today. For this reason, scholars have highlighted a temporal validity problem[183,184] and have questioned whether empirical findings in this space might have an expiration date[185]. Moreover, research suggests that people's general attitudes towards AI are evolving as these technologies become ever more embedded in daily life[186,187]. Future research should use longitudinal methods and developmental approaches to explore how trust in AI is changing over time, what predicts these changes in trust, and whether longitudinal changes vary across individuals and cultures.

Finally, and perhaps most importantly, the study of trust in AI must be placed in context. Fully understanding the behavioural science of trust in AI requires considering interdisciplinary and critical perspectives on what trust does. Considering trust in isolation from the social and political context in

which it occurs risks unintentional ethics-washing by reinforcing a discursive space in which talking about how to increase trust in AI implicitly supports the idea that a world of algorithms is desirable in the first place. Understanding trust in AI requires studying not only how people think, but also reflecting on the kind of world that trust in AI serves to create.

Table 1. Selected scales used to measure trust in AI.

Scale name	**Number of items**	**Items relating to performance and morality**	**Subscales or subdimensions**	**Example items**	**Internal consistency[a]**	**Selected other scale validation[b]**
System trustworthiness scale[188]	15	No[c]	Performance	"The system performs the task accurately."	α = 0.87 - 0.95	Reliability (Cronbach's alpha) Structural validity (EFA; CFA) Construct validity (convergent validity; divergent validity; criterion validity through experimental manipulation)
			Purpose	"The system is programmed specifically to complete this task"	α = 0.80 - 0.90	
			Process	"I understand how the system is supposed to work."	α = 0.82 - 0.92	
			-	-	Overall: Not assessed	
Human–computer trust scale[189]	12	Yes	Benevolence	"I believe that (—) will act in my best interest"	α = 0.83 - 0.84	Reliability (Cronbach's alpha; composite reliability) Structural validity (partial least square structural equation modelling)
			Competence	"I think that (—) is competent and effective in (—)"	α = 0.84 - 0.88	
			Perceived risk	"I believe that there could be negative consequences when using (—)"	α = 0.86 - 0.90	
			Trust	"I can trust the information presented to me by (—)"	α = 0.84 - 0.89	
			-	-	Overall: Not assessed	

Trust in automation scale[5]	12	Yes	None	"The system is deceptive" "I am confident in the system"	Not assessed	Structural validity (PCA; cluster analysis)
Trust in automation questionnaire[190]	19	No	Reliability/ competence	"The system is capable of interpreting situations correctly"	$\omega t = 0.92$	Reliability (Omega; Cronbach's alpha; Spearman-Brown coefficient) Structural validity (EFA with parallel analysis; model fit evaluated with CFA-type indices) Construct validity (convergent validity with single trust item; criterion validity through experimental manipulation; predictive validity)
			Understandability/ predictability	"I was able to understand why things happened"	$\omega t = 0.81$	
			Familiarity	"I have already used similar systems"	$r_{SB} = 0.83$	
			Intention of developers	"The developers are trustworthy"	$r_{SB} = 0.79$	
			Propensity to trust	"I rather trust a system than I mistrust it"	$\omega t = 0.78$	
			[General] Trust in automation	"I trust the system"	$\alpha = 0.63–0.85$	
			-	-	Overall : Not assessed	
Trust attitude measurement instrument[191]	16	No	Understandability	"I understand how the AI system works and I feel confident I will be able to use it in the future."	$\alpha = 0.79$	Reliability (Cronbach's alpha; omega; test-retest reliability; composite reliability) Structural validity (CFA) Construct validity (content validity through experts and content validity index;
			Technical competence	"The AI system uses appropriate methods to get results based on the information I input"	$\alpha = 0.84$	
			Reliability	"The AI system consistently provides the results it is expected to produce."	$\alpha = 0.72$	
			Helpfulness	"When I need help, the AI system responds to my needs effectively and responsively"	$\alpha = 0.87$	

			Personal attachment	"I like using the AI system because it suits me, and always want to use it"	$\alpha = 0.84$	
			User autonomy	"I feel in control when operating the various functions and features of the AI system"	$\alpha = 0.83$	
			Faith	"When I am unsure about the AI system's result, I believe in the AI system rather than myself."	$\alpha = 0.83$	
			Institutional credibility	"I feel assured using the AI system because it is made by a reputable institution and therefore already went through a credible regulation ."	$\alpha = 0.91$	
			-	-	Overall : $\alpha = 0.95$	
Human–computer trust[192]	25	No	Reliability	"I can rely on the system to function properly."	$\alpha = 0.85$	Reliability (Cronbach's alpha) Structural validity (PCA) Construct validity (content validity through nominal group technique) Other (inter-rater reliability)
			Technical competence	"The system uses appropriate methods to reach decisions"	$\alpha = 0.74$	
			Understandability	"It is easy to follow what the system does"	$\alpha = 0.84$	
			Faith	"I believe advice from the system even when I don' t know for certain that it is correct"	$\alpha = 0.88$	
			Personal attachment	"I like using the system for decision making."	$\alpha = 0.90$	
			-	-	Overall : $\alpha = 0.94$	
Multi-dimensional measure of trust[4]	16	Yes	Performance dimension: reliability	"Can count on"	$\alpha = 0.92$	Reliability (Cronbach's alpha) Structural validity (PCA)
			Performance	"Skilled"	$\alpha = 0.92$	

			dimension: capability			Construct validity (Content validity through sorting task; Criterion validity through experimental manipulation)
			Moral dimension: sincerity	"Authentic"	$\alpha = 0.79$	
			Moral dimension: ethicality	"Has integrity"	$\alpha = 0.81$	
			-		Overall : Not assessed	
Short trust in automation scale[193]	3	No	No	"I am confident in the AI assistant" "The AI assistant is reliable"	Overall : $\alpha = 0.97$	Reliability (Cronbach's alpha) Construct validity (Convergent validity with other scales; criterion validity through experimental manipulation; predictive validity)
Trust perception scale–human robot interaction[194], [d]	40	Yes	No	"What % of the time will this robot:" ... "Function successfully" " ... Protect people"	Not assessed	Structural validity (PCA) Construct validity (content validity through experts and content validity ratio; convergent validity with other scales; criterion validity through experimental manipulation)
AI trust score[195]	4	No	No	"[The AI feature] will help me do my job more efficiently and effectively." "I understand how and when to use [the AI feature]."	Overall : $\alpha = 0.70$	Reliability (Cronbach's alpha) Structural validity (EFA) Construct validity (convergent validity; content validity through qualitative data; predictive validity)
Human-generative artificial intelligence trust scale[196]	9	Yes	Benevolence	"I believe that artificial intelligence (such as ChatGPT, ERNIE Bo, etc.) acts in my best interest"	CR = 0.81	Reliability (Cronbach's alpha; composite reliability) Structural Validity

			Competence	"I believe that artificial intelligence (...) is capable of effectively supporting my learning process"	CR = .81	(ESEM; CFA) Construct Validity (convergent validity; criterion validity; content validity through translations)
			Reciprocity	"If I use artificial intelligence (...), I believe I can fully rely on it"	CR = .81	
			-	-	Overall: α = 0.89; CR = .93	
Trust in AI scale[197]	30	Yes	Global Trust	"I trust %system%."	α = 0.92 - 0.94	Reliability (Cronbach's alpha) Structural validity (EFA; CFA) Construct validity (convergent validity) Other (measurement invariance)
			Vigilance	"I am careful when using %system%."	α = 0.90 - 0.91	
			Unbiasedness	"%system% was designed to be responsible."	α = 0.79 - 0.90	
			Integrity	"%system% is sincere."	α = 0.91 - 0.93	
			Transparency	"The inner workings of %system% are comprehensible."	α = 0.84 - 0.86	
			Ability	"%system% is capable."	α = 0.95 - 0.96	
			-	-	Overall: Not assessed	

Note. α: Cronbach's alpha; r_{SB}: Spearman-Brown coefficient; ωt: Omega total; ωr: Revelle's omega; CR: Composite Reliability; EFA: Exploratory Factor Analysis; CFA: Confirmatory Factor Analysis; ESEM: Exploratory Structural Equation Modelling; PCA: Principal Components Analysis

[a] Although we report internal consistency measurements taken from the cited publications, it is important to place these values within our broader arguments about the agent-specificity, multi-dimensionality, and even temporal validity of trust in AI. Some scales have been developed with a specific AI type in mind (for example, chatbots) and might not show the same internal consistency for other types of AI, and reliability estimates for scales with only performance items may be higher than those with moral trust items.

[b] There are different methods and approaches to scale validation and similar terms can be used in different ways (for example, convergent validity typically refers to whether scores on a scale converge with other established scales, but in some work it can instead be used to refer to whether the individual subscales are correlated. We present a selection of common validation, approaches used to highlight the diversity in this space, not to critique or compare scales.

[c] The system trustworthiness scale follows the adaptation of the tripartite trust model based on ability, benevolence, and integrity[25] to distinguish trust in AI based on performance, purpose, and process[62]. However, the items in the purpose subscale here are more focused on whether the machine is designed for a specific purpose by its developers, not whether the system necessarily has good intentions or morality. A system can be designed with unethical intentions in mind and used in that way, but that does not mean the system would be (perceived as) benevolent or having a good purpose.

[d] This scale includes a short 14-item version which can be used to measure change in trust over time or in an assessment with multiple trials, which focuses on the functional capabilities of the system.

Table 2. Implications for understanding, building, and regulating AI.

Principle	Meaning	Implications for Researchers	Implications for Developers	Implications for Policymakers
Trust in AI is inferred	Trust is a psychologically inferred process based on social cognition; it does not lie in computer code.	Theory and measurement should distinguish between actual trustworthiness and perceived trustworthiness.	Designing trustworthy systems does not guarantee that they will be perceived as trustworthy by users. Systems should be designed to be trustworthy, not trusted.	Regulation should not assume that certified or compliant AI will automatically be trusted by the public.
Trustworthiness, trust, and trusting behaviour are distinct	Trust in AI can manifest in conceptually and empirically different ways.	Theory and measurement should distinguish between trustworthiness, trust, and trusting behaviour.	User trust might not translate into use, and use might occur without trust.	AI use should not be taken as evidence of public trust or acceptance.
Trust in AI is multidimensional	Trust in AI depends on perceptions of both performance and morality.	Research should explore performance and moral dimensions of trust separately.	Improving performance alone might not increase moral trust in an AI system.	Governance frameworks should address ethical concerns, not just technical performance.
Trust in AI is agent-specific	People trust different AI systems differently depending on context, purpose, and domain.	Findings about trust in 'AI' might not generalise across different systems and applications.	Trust cues vary across specific domains of AI application.	Regulation should take domain differences into account rather than treating all AI systems as the same.
Trust in AI is individually variable	Trust in AI varies across individuals and cultures.	Research should account for and explore individual and cultural differences when studying trust in AI.	AI designers should anticipate diverse trust responses rather than a single user profile.	Regulation should account for the fact that trust in AI will not be uniform across populations.
Trust in AI is strategically motivated	People might trust or rely on AI based on strategic motivations, incentives, and goals.	Research should explore how strategic motivations shape trust and reliance on AI systems.	Developers should anticipate that users might over-rely on AI in situations where reduced responsibility or effort are desirable.	Regulation should address strategic misuse or overreliance.

Figure captions

Figure 1: Trustworthiness, trust, and trusting behaviour. The relationship between trustworthiness, trust, and trusting behaviour as typically understood based on the tripartite model of trust[25] and its adaptation to trust in AI[48,62].

Figure 2: The meaning of 'trust in AI'. A) Trustworthiness can refer an actual assessment of whether the system is deserving of trust[1,9,21,44] or a perception of whether the system deserves trust[12]. B) Perceived trustworthiness can be driven by expectations of both performance (for example, reliability or capability) and morality (for example, ethicality, sincerity or benevolence)[3,13,84], which can also be called bases of trust. C) Trust can be an attitude reflecting a willingness to rely on another in a situation characterised by uncertainty (attitudinal trust)[11,12,] measured by explicit questions like "I can trust the system"[63]. D) Trust can be an attitude reflecting or including the perception of the system's capabilities that would enable someone to rely on it in a situation of uncertainty, measured by questions like "The system is reliable" and "The system is ethical"[3,63]. E) Trust can be used broadly to refer to all the psychological aspects of trust, including an assessment of the system, a self-reported willingness to rely on the system, and a behaviour, distinct only from actual trustworthiness[4,34,45].

Box 1. Trust terminology

Actual trustworthiness
Trustworthiness based on objective characteristics that make someone or something worthy of trust.

Perceived trustworthiness
Trustworthiness based on the perception of characteristics that make someone or something worthy of trust.

Attitudinal trust
Trust as an attitude that reflects a willingness to rely on someone or something in a situation characterised by uncertainty.

Cognitive trust
Trust as rationally concluded based on evidence of someone's or something's capacities.

Affective trust
Trust based on an emotional response to someone or something.

Performance trust
Trust in the reliability, ability, and performance of someone or something.

Moral trust
Trust in someone or something to be ethical, sincere, genuine and authentic, follow moral norms and principles, and seek to do good.

Calibrated trust
Trust that is appropriate to the degree to which the object of trust is trustworthy.

Trusting behaviour
Behaviour that reflects a trusting attitude towards someone or something.

Notes

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Psychology thanks Shenghua Luan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

References

1. Li, B. *et al.* Trustworthy AI: from principles to practices. *ACM Comput. Surv.* **55**, 1–46 (2023).
2. Kaplan, A. D., Kessler, T. T., Brill, J. C. & Hancock, P. A. Trust in artificial intelligence: meta-analytic findings. *Hum. Factors* **65**, 337–359 (2023).
3. Lalot, F. & Bertram, A.-M. When the bot walks the talk: Investigating the foundations of trust in an artificial intelligence (AI) chatbot. *J. Exp. Psychol. Gen.* **154**, 533–551 (2025).
4. Malle, B. F. & Ullman, D. A multidimensional conception and measure of human-robot trust. in *Trust in Human-Robot Interaction* 3–25 (Elsevier, 2021). doi:10.1016/B978-0-12-819472-0.00001-0.
5. Jian, J.-Y., Bisantz, A. M. & Drury, C. G. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *Int. J. Cogn. Ergon.* **4**, 53–71 (2000).
6. Jermutus, E., Kneale, D., Thomas, J. & Michie, S. Influences on user trust in healthcare artificial intelligence: A systematic review. *Wellcome Open Res.* **7**, 65 (2022).
7. Bach, T. A., Khan, A., Hallock, H., Beltrão, G. & Sousa, S. A systematic literature review of user trust in AI-enabled systems: an HCI perspective. *Int. J. Human–Computer Interact.* **40**, 1251–1266 (2024).
8. Floridi, L. & Cowls, J. A unified framework of five principles for AI in society. *Harv. Data Sci. Rev.* https://doi.org/10.1162/99608f92.8cd550d1 (2019) doi:10.1162/99608f92.8cd550d1.
9. Thiebes, S., Lins, S. & Sunyaev, A. Trustworthy artificial intelligence. *Electron. Mark.* **31**, 447–464 (2021).
10. European Commision. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
11. Afroogh, S., Akbari, A., Malone, E., Kargar, M. & Alambeigi, H. Trust in AI: progress, challenges, and future directions. *Humanit. Soc. Sci. Commun.* **11**, 1568 (2024).
12. Dang, Q. & Li, G. Unveiling trust in AI: the interplay of antecedents, consequences, and cultural dynamics. *AI Soc.* https://doi.org/10.1007/s00146-025-02477-6 (2025) doi:10.1007/s00146-025-02477-6.
13. Li, Y., Wu, B., Huang, Y. & Luan, S. Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. *Front. Psychol.* **15**, 1382693 (2024).

14. Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. M. & Tielman, M. L. A systematic review on fostering appropriate trust in human-AI interaction: trends, opportunities and challenges. *ACM J. Responsible Comput.* **1**, 1–45 (2024).
15. Hoff, K. A. & Bashir, M. Trust in automation: integrating empirical evidence on factors that influence trust. *Hum. Factors J. Hum. Factors Ergon. Soc.* **57**, 407–434 (2015).
16. Burton, J. W., Stein, M. & Jensen, T. B. A systematic review of algorithm aversion in augmented decision making. *J. Behav. Decis. Mak.* **33**, 220–239 (2020).
17. Mahmud, H., Islam, A. K. M. N., Ahmed, S. I. & Smolander, K. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technol. Forecast. Soc. Change* **175**, 121390 (2022).
18. Köbis, N., Bonnefon, J.-F. & Rahwan, I. Bad machines corrupt good morals. *Nat. Hum. Behav.* **5**, 679–685 (2021).
19. Bonnefon, J.-F., Rahwan, I. & Shariff, A. The moral psychology of artificial intelligence. *Annu. Rev. Psychol.* **75**, 653–675 (2024).
20. Marocco, S., Barbieri, B. & Talamo, A. Exploring facilitators and barriers to managers' adoption of AI-based systems in decision making: a systematic review. *AI* **5**, 2538–2567 (2024).
21. Montealegre-López, N. Exploring the role of trust in AI-driven decision-making: a systematic literature review. *Manag. Rev. Q.* https://doi.org/10.1007/s11301-025-00526-4 (2025) doi:10.1007/s11301-025-00526-4.
22. Jing, P., Xu, G., Chen, Y., Shi, Y. & Zhan, F. The determinants behind the acceptance of autonomous vehicles: a systematic review. *Sustainability* **12**, 1719 (2020).
23. Rahwan, I. *et al.* Machine behaviour. *Nature* **568**, 477–486 (2019).
24. Rousseau, D., Sitkin, S., Burt, R. & Camerer, C. Not so Different after All: A Cross-Discipline View of Trust. *Acad. Manage. Rev.* **23**, 393–404 (1998).
25. Mayer, R. C., Davis, J. H. & Schoorman, F. D. An integrative model of organizational trust. *Acad. Manage. Rev.* **20**, 709–734 (1995).
26. Lee, J. D. & See, K. A. Trust in automation: designing for appropriate reliance. *Hum. Factors J. Hum. Factors Ergon. Soc.* **46**, 50–80 (2004).
27. Metzinger, T. EU guidelines: ethics washing made in Europe. *Der Tagesspiegel Online* (2019).
28. Friedman, B., Khan, P. H. & Howe, D. C. Trust online. *Commun. ACM* **43**, 34–40 (2000).
29. Baier, A. Trust and Antitrust. *Ethics* **96**, 231–260 (1986).
30. Hawley, K. Trust, distrust and commitment. *Noûs* **48**, 1–20 (2014).
31. Viehoff, J. Making trust safe for AI? Non-agential trust as a conceptual engineering problem. *Philos. Technol.* **36**, 64 (2023).
32. Ryan, M. In AI we trust: ethics, artificial intelligence, and reliability. *Sci. Eng. Ethics* **26**, 2749–2767 (2020).
33. Freiman, O. Making sense of the conceptual nonsense 'trustworthy AI'. *AI Ethics* **3**, 1351–1360 (2023).
34. Zanotti, G. AI systems should be trustworthy, not trusted. *AI Soc.* https://doi.org/10.1007/s00146-025-02728-6 (2025) doi:10.1007/s00146-025-02728-6.

35. Nass, C. & Lee, K. M. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *J. Exp. Psychol. Appl.* **7**, 171–181 (2001).
36. Nass, C., Steuer, J. & Tauber, E. R. Computers are social actors. in *Conference companion on Human factors in computing systems - CHI '94* 204 (ACM Press, Boston, Massachusetts, United States, 1994). doi:10.1145/259963.260288.
37. Reeves, B. & Nass, C. I. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. xiv, 305 (Cambridge University Press, New York, NY, US, 1996).
38. Nass, C. & Moon, Y. Machines and mindlessness: social responses to computers. *J. Soc. Issues* **56**, 81–103 (2000).
39. Nass, C., Moon, Y., Fogg, B. J., Reeves, B. & Dryer, D. C. Can computer personalities be human personalities? *Int. J. Hum.-Comput. Stud.* **43**, 223–239 (1995).
40. Epley, N., Waytz, A. & Cacioppo, J. T. On seeing human: A three-factor theory of anthropomorphism. *Psychol. Rev.* **114**, 864–886 (2007).
41. Bocian, K., Gonidis, L. & Everett, J. A. C. Moral conformity in a digital world: Human and nonhuman agents as a source of social pressure for judgments of moral character. *PLOS ONE* **19**, e0298293 (2024).
42. Landes, E., Claessens, S. & Everett, J. A. C. Trust in AI is not a conceptual confusion. Preprint at https://philpapers.org/rec/LANTIA-8 (2026).
43. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **1**, 389–399 (2019).
44. Liu, H. *et al.* Trustworthy AI: a computational perspective. *ACM Trans. Intell. Syst. Technol.* **14**, 1–59 (2023).
45. Zerilli, J., Bhatt, U. & Weller, A. How transparency modulates trust in artificial intelligence. *Patterns* **3**, 100455 (2022).
46. Mayer, R. C., Davis, J. H. & Schoorman, F. D. An integrative model of organizational trust. *Acad. Manage. Rev.* **20**, 709 (1995).
47. Schlicker, N. & Langer, M. Towards warranted trust: a model on the relation between actual and perceived system trustworthiness. in *Proceedings of Mensch und Computer 2021* 325–329 (Association for Computing Machinery, New York, NY, USA, 2021). doi:10.1145/3473856.3474018.
48. Schlicker, N. *et al.* How do we assess the trustworthiness of AI? introducing the trustworthiness assessment model (TrAM). *Comput. Hum. Behav.* **170**, 108671 (2025).
49. Brambilla, M., Sacchi, S., Rusconi, P. & Goodwin, G. P. The primacy of morality in impression development: theory, research, and future directions. in *Advances in Experimental Social Psychology* (ed. Gawronski, B.) vol. 64 187–262 (Academic Press, 2021).
50. Cuddy, A. J. C., Fiske, S. T. & Glick, P. Warmth and competence as universal dimensions of social perception: the stereotype content model and the BIAS map. in *Advances in Experimental Social Psychology* vol. 40 61–149 (Academic Press, 2008).
51. Fiske, S. T., Cuddy, A. J. C. & Glick, P. Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* **11**, 77–83 (2007).

52. Willis, J. & Todorov, A. First impressions: making up your mind after a 100-ms exposure to a face. *Psychol. Sci.* **17**, 592–598 (2006).
53. Dion, K., Berscheid, E. & Walster, E. What is beautiful is good. *J. Pers. Soc. Psychol.* **24**, 285–290 (1972).
54. McKee, K. R., Bai, X. & Fiske, S. T. Humans perceive warmth and competence in artificial intelligence. *iScience* **26**, 107256 (2023).
55. Surdel, N. *et al.* Judging robot ability: how people form implicit and explicit impressions of robot competence. *J. Exp. Psychol. Gen.* **153**, 1309–1335 (2024).
56. Papenmeier, A., Kern, D., Hienert, D., Kammerer, Y. & Seifert, C. How Accurate Does It Feel? – Human Perception of Different Types of Classification Mistakes. in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* 1–13 (Association for Computing Machinery, New York, NY, USA, 2022). doi:10.1145/3491102.3501915.
57. Dwivedi, R. *et al.* Explainable AI (XAI): core ideas, techniques, and solutions. *ACM Comput. Surv.* **55**, 1–33 (2023).
58. Nussberger, A.-M., Luo, L., Celis, L. E. & Crockett, M. J. Public attitudes value interpretability but prioritize accuracy in artificial intelligence. *Nat. Commun.* **13**, 5821 (2022).
59. Wright, J. L., Chen, J. Y. C. & Lakhmani, S. G. Agent transparency and reliability in human–robot interaction: the influence on user confidence and perceived reliability. *IEEE Trans. Hum.-Mach. Syst.* **50**, 254–263 (2020).
60. Schmidt, P., Biessmann, F. & Teubner, T. Transparency and trust in artificial intelligence systems. *J. Decis. Syst.* **29**, 260–278 (2020).
61. Leichtmann, B., Hinterreiter, A., Humer, C., Streit, M. & Mara, M. Explainable Artificial Intelligence Improves Human Decision-Making: Results from a Mushroom Picking Experiment at a Public Art Festival. *Int. J. Human–Computer Interact.* **40**, 4787–4804 (2024).
62. Lee, J. D. & See, K. A. Trust in automation: designing for appropriate reliance. *Hum. Factors* **46**, 50–80 (2004).
63. McAllister, D. J. Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *Acad. Manage. J.* **38**, 24–59 (1995).
64. Nannestad, P. What Have We Learned About Generalized Trust, If Anything? *Annu. Rev. Polit. Sci.* **11**, 413–436 (2008).
65. Vereschak, O., Bailly, G. & Caramiaux, B. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proc ACM Hum-Comput Interact* **5**, 327:1-327:39 (2021).
66. Rice, S. Examining single- and multiple-process theories of trust in automation. *J. Gen. Psychol.* **136**, 303–322 (2009).
67. Xie, Y., Bodala, I. P., Ong, D. C., Hsu, D. & Soh, H. Robot capability and intention in trust-based decisions across tasks. in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* 39–47 (IEEE, Daegu, Korea (South), 2019). doi:10.1109/HRI.2019.8673084.
68. Claessens, S. *et al.* Trust in artificial moral advisors across cultures. Preprint at (2026).
69. Costello, T. H., Pennycook, G. & Rand, D. G. Durably reducing conspiracy beliefs through dialogues with AI. *Science* **385**, 1–12 (2024).

70. Lin, H. *et al.* Persuading voters using human–artificial intelligence dialogues. *Nature* **648**, 394–401 (2025).
71. Kelley, T. L. *Interpretation of Educational Measurements*. (World Book Company, 1927).
72. Gonzalez, O., MacKinnon, D. P. & Muniz, F. B. Extrinsic Convergent Validity Evidence to Prevent Jingle and Jangle Fallacies. *Multivar. Behav. Res.* **56**, 3–19 (2021).
73. Marsh, H. W. *et al.* The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *J. Educ. Psychol.* **111**, 331–353 (2019).
74. Berg, J., Dickhaut, J. & McCabe, K. Trust, reciprocity, and social history. *Games Econ. Behav.* **10**, 122–142 (1995).
75. Ishowo-Oloko, F. *et al.* Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nat. Mach. Intell.* **1**, 517–521 (2019).
76. Makovi, K., Sargsyan, A., Li, W., Bonnefon, J.-F. & Rahwan, T. Trust within human-machine collectives depends on the perceived consensus about cooperative norms. *Nat. Commun.* **14**, 3108 (2023).
77. von Schenk, A., Klockmann, V. & Köbis, N. Social Preferences Toward Humans and Machines: A Systematic Experiment on the Role of Machine Payoffs. *Perspect. Psychol. Sci.* **20**, 165–181 (2025).
78. Ajzen, I. & Fishbein, M. A Bayesian analysis of attribution processes. *Psychol. Bull.* **82**, 261–277 (1975).
79. Ajzen, I. & Fishbein, M. *Understanding Attitudes and Predicting Social Behavior*. (Prentice-Hall, Englewood Cliffs, NJ, 2002).
80. Visser, R., Peters, T. M., Scharlau, I. & Hammer, B. Trust, distrust, and appropriate reliance in (X)AI: A conceptual clarification of user trust and survey of its empirical evaluation. *Cogn. Syst. Res.* **91**, 101357 (2025).
81. Bhaskar, P., Misra, P. & Chopra, G. Shall I use ChatGPT? A study on perceived trust and perceived risk towards ChatGPT usage by teachers at higher education institutions. *Int. J. Inf. Learn. Technol.* **41**, 428–447 (2024).
82. Dekkal, M., Arcand, M., Prom Tep, S., Rajaobelina, L. & Ricard, L. Factors affecting user trust and intention in adopting chatbots: the moderating role of technology anxiety in insurtech. *J. Financ. Serv. Mark.* **29**, 699–728 (2024).
83. Shi, S., Gong, Y. & Gursoy, D. Antecedents of trust and adoption intention toward artificially intelligent recommendation systems in travel planning: a heuristic–systematic model. *J. Travel Res.* **60**, 1714–1734 (2021).
84. Landes, E., Francis, K. B. & Everett, J. A. C. People defer to AI moral advice, but not blindly. *Cognition* **272**, 106504 (2026).
85. Hancock, P. A. *et al.* A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Factors J. Hum. Factors Ergon. Soc.* **53**, 517–527 (2011).
86. Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **144**, 114–126 (2015).
87. Madhavan, P., Wiegmann, D. A. & Lacson, F. C. Automation failures on tasks easily performed by operators undermine trust in automated aids. *Hum. Factors J. Hum. Factors Ergon. Soc.* **48**, 241–256 (2006).

88. Rosendorf, O., Smetana, M. & Vranka, M. Autonomous weapons and ethical judgments: experimental evidence on attitudes toward the military use of "killer robots". *Peace Confl. J. Peace Psychol.* **28**, 177–183 (2022).
89. Barber, B. *The Logic and Limits of Trust*. (Rutgers University Press, New Brunswick, N.J, 1983).
90. Cook, J. & Wall, T. New work attitude measures of trust, organizational commitment and personal need non-fulfilment. *J. Occup. Psychol.* **53**, 39–52 (1980).
91. Deutsch, M. The effect of motivational orientation upon trust and suspicion. *Hum. Relat.* **13**, 123–139 (1960).
92. Hovland, C. I., Janis, I. L. & Kelley, H. H. *Communication and Persuasion; Psychological Studies of Opinion Change*. xii, 315 (Yale University Press, New Haven, CT, US, 1953).
93. Glikson, E. & Woolley, A. W. Human Trust in Artificial Intelligence: Review of Empirical Research. *Acad. Manag. Ann.* **14**, 627–660 (2020).
94. Schaefer, K. E., Chen, J. Y. C., Szalma, J. L. & Hancock, P. A. A Meta-Analysis of Factors Influencing the Development of Trust in Automation. *Hum. Factors* https://doi.org/10.1177/0018720816634228 (2016) doi:10.1177/0018720816634228.
95. Hicks, M. T., Humphries, J. & Slater, J. ChatGPT is bullshit. *Ethics Inf. Technol.* **26**, 38 (2024).
96. Phan, T. A. & Bui, V. D. AI with a heart: how perceived authenticity and warmth shape trust in healthcare chatbots. *J. Mark. Commun.* 1–21 (2025) doi:10.1080/13527266.2025.2508887.
97. Myers, S. & Everett, J. A. C. Is an intelligent machine a moral machine? *Exp. Philos.* (In Press).
98. Armstrong, S. General purpose intelligence: Arguing the orthogonality thesis. *Anal. Metaphys.* 68–84 (2013).
99. Bostrom, N. The superintelligent will: motivation and instrumental rationality in advanced artificial agents. *Minds Mach.* **22**, 71–85 (2012).
100. Claessens, S. & Everett, J. A. C. Trust In artificial intelligence Is agent-specific and multidimensional. Preprint at https://osf.io/preprints/psyarxiv/4y6xw_v1 (2026).
101. Langer, M. *et al.* What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif. Intell.* **296**, 103473 (2021).
102. de Visser, E. J. *et al.* Almost human: anthropomorphism increases trust resilience in cognitive agents. *J. Exp. Psychol. Appl.* **22**, 331–349 (2016).
103. Gomes, S., Lopes, J. M. & Nogueira, E. Anthropomorphism in artificial intelligence: a game-changer for brand marketing. *Future Bus. J.* **11**, 2 (2025).
104. Gillespie, N., Lockey, S., Ward, T., Macdade, A. & Hassed, G. *Trust, Attitudes and Use of Artificial Intelligence: A Global Study 2025*. (2025) doi:10.26188/28822919.
105. Bigman, Y. E. & Gray, K. People are averse to machines making moral decisions. *Cognition* **181**, 21–34 (2018).
106. Myers, S. & Everett, J. A. C. People expect artificial moral advisors to be more utilitarian and distrust utilitarian moral advisors. *Cognition* **256**, 106028 (2025).

107. Dong, M., Conway, J. R., Bonnefon, J.-F., Shariff, A. & Rahwan, I. Fears about artificial intelligence across 20 countries and six domains of application. *Am. Psychol.* https://doi.org/10.1037/amp0001454 (2024) doi:10.1037/amp0001454.
108. Westjohn, S. A., Magnusson, P., Franke, G. R. & Peng, Y. Trust propensity across cultures: the role of collectivism. *J. Int. Mark.* **30**, 1–17 (2022).
109. Scholz, D. D., Kraus, J. & Miller, L. Measuring the propensity to trust in automated technology: examining similarities to dispositional trust in other humans and validation of the PTT-a scale. *Int. J. Human–Computer Interact.* **41**, 970–993 (2025).
110. Montag, C., Becker, B. & Li, B. J. On trust in humans and trust in artificial intelligence: a study with samples from Singapore and Germany extending recent research. *Comput. Hum. Behav. Artif. Hum.* **2**, 100070 (2024).
111. Riedl, R. Is trust in artificial intelligence systems related to user personality? review of empirical evidence and future research directions. *Electron. Mark.* **32**, 2021–2051 (2022).
112. Böckle, M., Yeboah-Antwi, K. & Kouris, I. Can you trust the black box? The effect of personality traits on trust in AI-enabled user interfaces. in *Artificial Intelligence in HCI* (eds Degen, H. & Ntoa, S.) 3–20 (Springer International Publishing, Cham, 2021). doi:10.1007/978-3-030-77772-2_1.
113. Aliasghari, P., Ghafurian, M., Nehaniv, C. L. & Dautenhahn, K. Effect of domestic trainee robots' errors on human teachers' trust. in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)* 81–88 (IEEE, Vancouver, BC, Canada, 2021). doi:10.1109/RO-MAN50785.2021.9515510.
114. Zhang, T. *et al.* Automated vehicle acceptance in China: social influence and initial trust are key determinants. *Transp. Res. Part C Emerg. Technol.* **112**, 220–233 (2020).
115. Müller, L., Mattke, J., Maier, C., Weitzel, T. & Graser, H. Chatbot acceptance: a latent profile analysis on individuals' trust in conversational agents. in *Proceedings of the 2019 on Computers and People Research Conference* 35–42 (ACM, Nashville TN USA, 2019). doi:10.1145/3322385.3322392.
116. Schandl, F., Fischer, P. & Hudecek, M. F. C. Predicting acceptance of autonomous shuttle buses by personality profiles: a latent profile analysis. *Transportation* **52**, 1015–1038 (2025).
117. Montag, C., Kraus, J., Baumann, M. & Rozgonjuk, D. The propensity to trust in (automated) technology mediates the links between technology self-efficacy and fear and acceptance of artificial intelligence. *Comput. Hum. Behav. Rep.* **11**, 100315 (2023).
118. Gillath, O. *et al.* Attachment and trust in artificial intelligence. *Comput. Hum. Behav.* **115**, 106607 (2021).
119. Gallimore, D., Lyons, J. B., Vo, T., Mahoney, S. & Wynne, K. T. Trusting Robocop: Gender-Based Effects on Trust of an Autonomous Robot. *Front. Psychol.* **10**, 482 (2019).
120. Heerink, M. Exploring the influence of age, gender, education and computer experience on robot acceptance by older adults. in *Proceedings of the 6th international conference on Human-robot interaction* 147–148 (ACM, Lausanne Switzerland, 2011). doi:10.1145/1957656.1957704.
121. Rahman, M. M., Babiker, A. & Ali, R. Motivation, concerns, and attitudes towards AI: differences by gender, age, and culture. in *Web Information Systems Engineering – WISE 2024* (eds Barhamgi, M.,

Wang, H. & Wang, X.) 375–391 (Springer Nature, Singapore, 2025). doi:10.1007/978-981-96-0573-6_28.
122. Booth, R. English-speaking countries more nervous about rise of AI, polls suggest. *The Guardian* (2025).
123. Liu, N. T. Y., Kirshner, S. N. & Lim, E. T. K. Is algorithm aversion WEIRD? A cross-country comparison of individual-differences and algorithm aversion. *J. Retail. Consum. Serv.* **72**, 103259 (2023).
124. Simon, M., Houghton, S. M. & Aquino, K. Cognitive biases, risk perception, and venture formation: How individuals decide to start companies. *J. Bus. Ventur.* **15**, 113–134 (2000).
125. Tetlock, P. E. Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychol. Rev.* **109**, 451–471 (2002).
126. Ditto, P. H., Pizarro, D. A. & Tannenbaum, D. Motivated moral reasoning. *Psychol. Learn. Motiv.* **50**, 307–338 (2009).
127. Haidt, J. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychol. Rev.* **108**, 814–834 (2001).
128. Parasuraman, R. & Manzey, D. H. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Hum. Factors* **52**, 381–410 (2010).
129. Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc. JAMIA* **19**, 121–127 (2012).
130. Klayman, J. Varieties of Confirmation Bias. in *Psychology of Learning and Motivation* vol. 32 385–418 (Academic Press, 1995).
131. Kuhn, D. & Lao, J. Effects of Evidence on Attitudes: Is Polarization the Norm? *Psychol. Sci.* **7**, 115–120 (1996).
132. Lord, C. G., Ross, L. & Lepper, M. R. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol.* **37**, 2098–2109 (1979).
133. Leung, E. & Urminsky, O. The narrow search effect and how broadening search promotes belief updating. *Proc. Natl. Acad. Sci.* **122**, e2408175122 (2025).
134. Rathje, S. *et al.* Sycophantic AI increases attitude extremity and overconfidence. Preprint at https://doi.org/10.31234/osf.io/vmyek_v1 (2025).
135. Claessens, S., Veitch, P. & Everett, J. A. C. Negative perceptions of outsourcing to artificial intelligence. *Comput. Hum. Behav.* **177**, 108894 (2026).
136. Brambilla, M., Rusconi, P., Sacchi, S. & Cherubini, P. Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *Eur. J. Soc. Psychol.* **41**, 135–143 (2011).
137. Melnikoff, D. E. & Bailey, A. H. Preferences for moral vs. immoral traits in others are conditional. *Proc. Natl. Acad. Sci.* **115**, (2018).
138. Bonnefon, J.-F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
139. Ng, Y.-L. Understanding passenger acceptance of autonomous vehicles through the prism of the trolley dilemma. *Int. J. Human–Computer Interact.* **40**, 2185–2194 (2024).
140. Miazek, K. & Bocian, K. When AI is fairer than humans: the role of egocentrism in moral and fairness judgments of AI and human decisions. *Comput. Hum. Behav. Rep.* **19**, 100719 (2025).

141. Köbis, N. *et al.* Delegation to artificial intelligence can increase dishonest behaviour. *Nature* **646**, 126–134 (2025).
142. Trope, Y. & Liberman, N. Construal-level theory of psychological distance. *Psychol. Rev.* **117**, 440–463 (2010).
143. King, T. C., Aggarwal, N., Taddeo, M. & Floridi, L. Artificial intelligence crime: an interdisciplinary analysis of foreseeable threats and solutions. *Sci. Eng. Ethics* **26**, 89–120 (2020).
144. Paharia, N., Kassam, K. S., Greene, J. D. & Bazerman, M. H. Dirty work, clean hands: the moral psychology of indirect agency. *Organ. Behav. Hum. Decis. Process.* **109**, 134–141 (2009).
145. Jha, K., Doshi, A., Patel, P. & Shah, M. A comprehensive review on automation in agriculture using artificial intelligence. *Artif. Intell. Agric.* **2**, 1–12 (2019).
146. Javaid, M., Haleem, A., Khan, I. H. & Suman, R. Understanding the potential applications of Artificial Intelligence in Agriculture Sector. *Adv. Agrochem* **2**, 15–30 (2023).
147. van Klompenburg, T., Kassahun, A. & Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **177**, 105709 (2020).
148. Abid, S. K. *et al.* Toward an Integrated Disaster Management Approach: How Artificial Intelligence Can Boost Disaster Management. *Sustainability* **13**, (2021).
149. Sun, W., Bocchini, P. & Davison, B. D. Applications of artificial intelligence for disaster management. *Nat. Hazards* **103**, 2631–2689 (2020).
150. Camps-Valls, G. *et al.* Artificial intelligence for modeling and understanding extreme weather and climate events. *Nat. Commun.* **16**, 1919 (2025).
151. Fergus, P. *et al.* Harnessing Artificial Intelligence for Wildlife Conservation. *Conservation* **4**, 685–702 (2024).
152. Isabelle, D. A., Westerlund, M., Isabelle, D. A. & Westerlund, M. A Review and Categorization of Artificial Intelligence-Based Opportunities in Wildlife, Ocean and Land Conservation. *Sustainability* **14**, (2022).
153. Serrano, D. R. *et al.* Artificial Intelligence (AI) Applications in Drug Discovery and Drug Delivery: Revolutionizing Personalized Medicine. *Pharmaceutics* **16**, 1328 (2024).
154. Yang, Z., Zeng, X., Zhao, Y. & Chen, R. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct. Target. Ther.* **8**, 115 (2023).
155. Kahr, P. & Bernstein, A. Enough With Trust! Why We Must Move Beyond a Convenient but Insufficient Concept in AI-Supported Decision-Making Research. (2026).
156. Klingbeil, A., Grützner, C. & Schreck, P. Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Comput. Hum. Behav.* **160**, 108352 (2024).
157. Holbrook, C., Holman, D., Clingo, J. & Wagner, A. R. Overtrust in AI Recommendations About Whether or Not to Kill: Evidence from Two Human-Robot Interaction Studies. *Sci. Rep.* **14**, 19751 (2024).
158. Buçinca, Z., Malaya, M. B. & Gajos, K. Z. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* **5**, 1–21 (2021).

159. Narayanan, A. & Kapoor, S. *AI Snake Oil*. (Princeton University Press, 2024).
160. Buiten, M. C. Towards Intelligent Regulation of Artificial Intelligence. *Eur. J. Risk Regul.* **10**, 41–59 (2019).
161. Woersdoerfer, M. Ten reasons why—the case for more and better AI regulation. *AI Ethics* **6**, 62 (2025).
162. Scharowski, N., Benk, M., Kühne, S. J., Wettstein, L. & Brühlmann, F. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* 248–260 (Association for Computing Machinery, New York, NY, USA, 2023). doi:10.1145/3593013.3593994.
163. Krüger, S. & Wilson, C. The problem with trust: on the discursive commodification of trust in AI. *AI Soc.* **38**, 1753–1761 (2023).
164. Schultz, M. D., Conti, L. G. & Seele, P. Digital ethicswashing: a systematic review and a process-perception-outcome framework. *AI Ethics* **5**, 805–818 (2025).
165. Seele, P. & Schultz, M. D. From Greenwashing to Machinewashing: A Model and Future Directions Derived from Reasoning by Analogy. *J. Bus. Ethics* **178**, 1063–1089 (2022).
166. Sedlakova, J. & Trachsel, M. Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? *Am. J. Bioeth.* **23**, 4–13 (2023).
167. Ryan, M. In AI we trust: Ethics, artificial intelligence, and reliability. *Sci. Eng. Ethics* **26**, 2749–2767 (2020).
168. Duenser, A. & Douglas, D. M. Whom to trust, how and why: untangling artificial intelligence ethics principles, trustworthiness, and trust. *IEEE Intell. Syst.* **38**, 19–26 (2023).
169. Rubel, A., Castro, C. & Pham, A. Agency Laundering and Information Technologies. *Ethical Theory Moral Pract.* **22**, 1017–1041 (2019).
170. Bryson, J. J. Robots should be slaves. in *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (ed. Wilks, Y.) 63–74 (John Benjamins Publishing Company, 2010).
171. Skitka, L. J., Mosier, K. L. & Burdick, M. Does automation bias decision-making? *Int. J. Hum.-Comput. Stud.* **51**, 991–1006 (1999).
172. Salatino, A., Prével, A., Caspar, E. & Lo Bue, S. Influence of AI behavior on human moral decisions, agency, and responsibility. *Sci. Rep.* **15**, 12329 (2025).
173. Landes, E. Is it a tastytaste or a greedgrab? The importance of label choice in language design. *Philos. Psychol.* **0**, 1–28 (2025).
174. Fischer, E. & Sytsma, J. Zombie intuitions. *Cognition* **215**, 104807 (2021).
175. Fischer, E. Conceptual control: on the feasibility of conceptual engineering. *Inquiry* **68**, 3043–3071 (2025).
176. Danaher, J. The threat of algocracy: reality, resistance and accommodation. *Philos. Technol.* **29**, 245–268 (2016).
177. Vallor, S. *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. (Oxford University press, New York (N.Y.), 2024).

178. Landes, E. & Everett, J. A. C. AI should develop human empathy, not replace It. in *Empathy and Artificial Intelligence: Challenges, Advances, and Ethical Considerations* (eds Perry, A. & Cameron, C. D.) (Cambridge University Press, In Press).
179. Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. (Profile books, London, 2019).
180. Birhane, A. Algorithmic injustice: a relational ethics approach. *Patterns* **2**, 100205 (2021).
181. Crawford, K. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. (Yale University Press, New Haven London, 2021).
182. Muldoon, J. & Wu, B. A. Artificial intelligence in the colonial matrix of power. *Philos. Technol.* **36**, 80 (2023).
183. Rahwan, I., Shariff, A. & Bonnefon, J.-F. The science fiction science method. *Nature* **644**, 51–58 (2025).
184. Munger, K. Temporal validity as meta-science. *Res. Polit.* **10**, 20531680231187271 (2023).
185. Messeri, L. & Crockett, M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).
186. Daly, S. J., Wiewiora, A. & Hearn, G. Shifting attitudes and trust in AI: Influences on organizational AI adoption. *Technol. Forecast. Soc. Change* **215**, 124108 (2025).
187. Modhvadia, R., Sippy, T., Field Reid, O. & Margetts, H. *'How Do People Feel about AI?'* https://attitudestoai.uk/ (2025).
188. Alarcon, G. M. *et al.* Development and Validation of the System Trustworthiness Scale. *Hum. Factors J. Hum. Factors Ergon. Soc.* **66**, 1893–1913 (2024).
189. Gulati, S., Sousa, S. & Lamas, D. Design, development and evaluation of a human-computer trust scale. *Behav. Inf. Technol.* **38**, 1004–1015 (2019).
190. Körber, M. Theoretical considerations and development of a questionnaire to measure trust in automation. in *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (eds Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T. & Fujita, Y.) 13–30 (Springer International Publishing, Cham, 2019). doi:10.1007/978-3-319-96074-6_2.
191. Larasati, R. Human and AI Trust: Trust Attitude Measurement Instrument. Preprint at https://doi.org/10.48550/ARXIV.2510.21535 (2025).
192. Madsen, M. & Gregor, S. Measuring Human-Computer Trust.
193. McGrath, M. J., Lack, O., Tisch, J. & Duenser, A. Measuring trust in artificial intelligence: validation of an established scale and its short form. *Front. Artif. Intell.* **8**, (2025).
194. Schaefer, K. E. Measuring trust in human robot interactions: development of the "Trust Perception Scale-HRI". in *Robust Intelligence and Trust in Autonomous Systems* (eds Mittu, R., Sofge, D., Wagner, A. & Lawless, W. F.) 191–218 (Springer US, Boston, MA, 2016). doi:10.1007/978-1-4899-7668-0_10.
195. Wang, J. & Moulden, A. AI Trust Score: A User-Centered Approach to Building, Designing, and Measuring the Success of Intelligent Workplace Features. in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* 1–7 (ACM, Yokohama Japan, 2021). doi:10.1145/3411763.3443452.

196. Wang, P. *et al.* A Validation of the Human-Generative Artificial Intelligence Trust Scale. *Int. J. Human–Computer Interact.* 1–14 (2025) doi:10.1080/10447318.2025.2542881.
197. Wischnewski, M., Doebler, P. & Krämer, N. Development and validation of the Trust in AI Scale (TAIS). Preprint at https://doi.org/10.31234/osf.io/eqa9y_v1 (2025).