



Beyond the machine: risk, fear, optimism and the foundations of public trust in AI

Steven David Pickering^{1,2} · Martin Ejnar Hansen² · Yosuke Sunahara³

Received: 7 July 2025 / Accepted: 29 July 2026
© The Author(s) 2026

Abstract

What drives public trust in artificial intelligence (AI)? This study examines the individual and institutional foundations of AI trust across two contrasting democracies: Japan and the United Kingdom. Drawing on original survey data ($N=3235$), we test a set of hypotheses derived from trust-transfer perspectives and self-efficacy research, covering institutional trust, AI self-efficacy, technological optimism, perceived societal threat, and job displacement anxiety. The results show that trust in AI is shaped by both psychological predispositions and broader beliefs about the trustworthiness of political and scientific institutions. Trust in government, university scientists, and other people consistently predicts AI trust in both countries, even when controlling for demographic and attitudinal variables. While optimism about AI's benefits increases trust in both contexts, fear of AI plays a stronger negative role in the UK. Unexpectedly, the belief that AI will replace one's job is positively associated with trust in Japan but unrelated in the UK. These findings highlight how national context shapes public confidence in emerging technologies and point to the importance of governance frameworks that foster informed capability and institutional legitimacy.

Keywords AI · Trust · Japan · UK · Governance

1 Introduction

Artificial Intelligence (AI) is no longer confined to laboratories or speculative fiction; it is now embedded in everyday decision-making across public and private life. From predictive policing (Schiff et al. 2025) and automated welfare assessments to AI-assisted diagnostics (e.g. Rojahn et al. 2023) and autonomous transport, governments and corporations alike are deploying these technologies to increase efficiency, cut costs and offer new forms of public service delivery (Robles and Mallinson 2025). Yet, as AI becomes more pervasive, so too do questions about its legitimacy and

public acceptance, as well as questions of AI governance (David et al. 2024). Much like other emerging technologies, the effective use of AI in democratic societies depends on public trust, and trust is by some seen as an indicator of good governance (Bouckaert and Van de Walle 2003). Without a foundational level of trust, citizens may resist or reject AI systems, even when these are demonstrably beneficial. It is increasingly recognised that digital technologies require a “social licence” to operate; an informal public approval grounded in perceptions of legitimacy, fairness and accountability (Cheung and Ho 2025a). AI is no exception. Recent global surveys highlight that trust is a precondition for sustained adoption of AI technologies, particularly when these are deployed in sensitive domains such as health, security or employment (Glikson and Woolley 2020; Dreksler et al. 2025; WEF/Ipsos 2024). Conversely, perceived risks, ranging from data misuse and algorithmic bias to opaque decision-making and economic displacement, can severely undermine public confidence and hinder adoption (Bao et al. 2022; Kreps et al. 2023). As such, understanding what drives or diminishes trust in AI is not merely an academic concern but a pressing policy priority.

✉ Steven David Pickering
s.d.pickering@uva.nl

Martin Ejnar Hansen
martin.hansen@brunel.ac.uk

Yosuke Sunahara
sunahara@people.kobe-u.ac.jp

¹ University of Amsterdam, Amsterdam, The Netherlands

² Brunel University London, London, UK

³ Kobe University, Kobe, Japan

Trust in AI is a relatively new concept, yet several working definitions have already gained traction. Drawing on earlier organisational trust theory, [Siau and Wang \(2018\)](#) conceptualise trust in AI as a combination of beliefs about competence, benevolence, integrity and predictability, together with a willingness to rely on an AI system in situations involving risk. This formulation emphasises both cognitive evaluations and behavioural intention, rather than treating trust as a static attribute of the technology alone. More recent literature further underscores that trust is relational and context-dependent, not a property of the technology in isolation ([Montag et al. 2023](#)). While early studies focused on system-level attributes, such as accuracy, explainability, and transparency, other work suggests that individual attitudes, media framings, and institutional settings also shape how people judge the trustworthiness of AI systems ([Cheung and Ho 2025a, b](#); [Nguyen 2023](#); [Nguyen and Hekman 2024](#); [Robles and Mallinson 2025](#)). Despite the growing body of empirical work, most research has concentrated on technical or cognitive factors influencing trust, such as system performance, personal familiarity, or perceived competence ([Montag et al. 2023](#); [Kreps et al. 2023](#)). Much less attention has been given to the broader political and institutional contexts in which AI is introduced. Yet, there are compelling reasons to believe that trust in AI is closely tied to trust in the institutions that develop, govern, and deploy it. Studies have found, for example, that citizens are more likely to support AI systems operated by universities, independent research bodies, or local governments than those managed by large corporations or national political authorities ([WEF/Ipsos 2024](#)). As [Robles and Mallinson \(2025\)](#) argue, the effectiveness of AI governance ultimately hinges not only on technological capacity but also on political legitimacy and institutional trustworthiness. This also echoes “trust-transfer” perspectives, which propose that confidence in AI may draw on pre-existing trust in the scientists, developers, or regulators behind it ([Goh and Ho 2024](#); [Montag et al. 2024](#)).

Social science scholarship has long established that citizens’ willingness to support policies is conditioned by their trust in the political system. As [Levi and Stoker \(2000\)](#) argue, trust is a relational concept: it involves an individual placing themselves in a position of vulnerability in relation to an institution that possesses the power to help or harm them. Yet, trust is far from unconditional, with *generalised* trust related to desirable outcomes (e.g., [Newton and Zmerli 2011](#); [Zmerli and Newton 2008](#); [Van der Meer and Ouattara 2019](#)) and *specific* trust based on the judgement of a particular situation informed by previous experience ([Levi and Stoker 2000](#)). Trust in politics, e.g., trust in government, and institutional trust, e.g., trust in public services, science, etc., are critical mediators of public acceptance for new technologies ([Splichal 2022](#)). In the case of AI, these forms of trust may serve as a kind of cognitive

shortcut: when individuals lack detailed knowledge about how an AI system works, they rely instead on their general trust in the actors who develop or regulate it ([Schiff et al. 2025](#)). In this way, trust in AI may function as a “spill-over” effect from more generalised trust in institutions, a dynamic broadly compatible with trust-transfer perspectives ([Goh and Ho 2024](#); [Montag et al. 2024](#)).

As such, we investigate what affects trust in AI, with a particular focus on the role of political and institutional trust. We do so by conducting a cross-national comparison between two advanced industrial democracies: the United Kingdom and Japan. Both countries have developed national AI strategies, invested in public–private partnerships, and promoted AI applications in fields such as healthcare, social welfare, and transport. At the same time, they differ in important respects: Japan has emphasised ethical guidelines and a precautionary approach in public-sector AI deployment ([METI 2021](#)), while the UK has leaned into principles of transparency and accountability ([Domagala 2021](#)). Importantly, both countries exhibit relatively low levels of public trust in national government: according to the OECD, only about one-quarter of respondents report high trust in government ([OECD 2021, 2023](#)). These features make them ideal cases for examining how different institutional trust environments may shape citizens’ confidence in AI. [Novozhilova et al. \(2024\)](#) also highlight the benefits of studying trust perceptions of AI across different societal contexts.

Empirically, we draw on original nationally representative surveys conducted in November 2024 in both countries. The surveys measure citizens’ trust in a variety of AI applications and capture levels of trust in core public institutions, including government, the police, media, and private firms. This enables us to assess how general political trust and sector-specific institutional trust relate to AI-specific trust. While earlier studies have documented public attitudes toward AI, few have explicitly linked these attitudes to political or institutional trust, and even fewer have done so using a cross-national survey design. By filling this gap, the article aims to advance understanding of AI as not just a technological challenge, but as a challenge to governance. The remainder of the article proceeds as follows: we first review the relevant literature on trust in AI and institutional trust. We then present five theoretically informed hypotheses, before describing the survey data, analytical approach, and findings. In the conclusion, we reflect on the implications for AI governance and democratic legitimacy.

2 Trust in AI

Trust in AI has emerged as an important new research frontier, reflecting both the promises and societal risks posed by its rapid proliferation. Foundational work on AI

trust conceptualises it as a multi-dimensional phenomenon encompassing beliefs about competence, benevolence, and integrity (Mayer et al. 1995), a framework later applied to automation and AI concepts (Lee and See 2004; Siau and Wang 2018). This conceptualisation aligns with general theories of trust in technology, distinguishing between trust based on rational assessments (such as performance or explainability) and more diffuse, relational forms of trust (Lee and See 2004; Cheung and Ho 2025a). Research has further shown that trust is indispensable for AI's societal uptake, as public acceptance hinges not only on technical efficacy but also on perceived fairness, transparency and accountability (Kreps et al. 2023; Robles 2023; Thiebes et al. 2021). Several studies also point to a “trust paradox”: citizens may express willingness to adopt or use AI applications despite having limited or ambivalent trust in them (Kreps et al. 2023). This paradox reflects a tension between calculated risk–benefit perceptions and underlying reservations, often shaped by concerns over privacy, bias, and control (Bao et al. 2022). This tension also underlines the centrality of perceived risk and uncertainty in shaping how citizens negotiate their trust in AI systems. These perceptions are shaped by how experts present views on AI (e.g., Neri and Cozman 2020) and by how media represent AI-related risks (e.g., Schwarz 2024; Schwarz and Unsel 2024); see Krieger et al. (2024) for a review of AI risk perceptions.

Empirical studies directly linking political trust to AI attitudes are still, understandably, scarce (see Liehner et al. 2023; Dreksler et al. 2025; Gillespie et al. 2025). Some related work on algorithmic governance indicates that trust in authorities can shape support for AI-driven policies. For instance, surveys in Europe have found that citizens with higher trust in government tend to be more supportive of, or less resistant to, automated decision-making by public agencies (Araujo et al. 2023). By analogy, we expect that political and institutional trust will be important antecedents of AI trust: people who generally trust their government, judiciary or regulatory bodies may extend that trust to AI applications sponsored by those bodies. Yet, we also expect there to be individual-level psychological and attitudinal factors, such as risk tolerance, self-efficacy, optimism, and fear, that independently shape trust in AI, alongside institutional influences.

Social sciences offer a useful framework as they distinguish different types of trust. Political trust usually refers to citizens' confidence in political institutions (government, parliament) and leaders, whereas institutional trust can extend to other bodies (judiciary, media, NGOs, companies). A long literature in democratic theory has documented worrying trends in political trust: for instance, a global analysis found that trust in representative institutions has generally declined in many democracies over recent decades (Norris 2011), fuelling talk of a “crisis of democracy”. These

declines vary by country, but the United Kingdom is among those where trust in government has fallen sharply (Curtice and Montagu 2022). Even in societies like Japan, where governmental authority has traditionally been highly valued, trust in the national government is surprisingly low: as mentioned above, only about 24% of Japanese respondents reported high or moderate trust in their central government in 2021 (OECD 2021). Such low political trust has broad implications, as it correlates with reduced compliance with policies and lower civic engagement.

Recent literature has stressed that trust in AI does not operate in isolation from broader societal trust. Indeed, political and institutional trust (confidence in actors such as governments, scientists, and regulators) plays a foundational role in shaping attitudes to AI (Schiff 2024). When citizens trust the institutions that develop, govern, or oversee AI, they are more likely to trust the technologies themselves (Robles and Mallinson 2025; Dreksler et al. 2025). Survey evidence indicates that universities and public-sector bodies tend to enjoy higher trust than private companies or political actors, and this translates into more positive perceptions of AI associated with these institutions (WEF/Ipsos 2024; Splichal 2022). Furthermore, trust in public science and governance is associated with greater willingness to accept AI in contexts perceived as uncertain or high risk (Cheung and Ho 2025b). These dynamics align with trust-transfer theory, which suggests that individuals' trust in AI systems may draw on trust in the human or institutional actors associated with their development and oversight (Goh and Ho 2024; Montag et al. 2023, 2024).

3 Hypotheses

Bringing together these theoretical insights, we argue that public trust in AI is shaped by a combination of institutional, psychological, and affective factors. Trust in AI does not emerge in a vacuum; rather, it reflects broader confidence in the institutions that develop, regulate and deploy these technologies. At the same time, individuals' personal capabilities and emotional responses toward AI, such as self-efficacy, optimism, or fear, play an important role in how trustworthy AI systems are perceived to be. Our expectation is that both the societal and individual levels jointly shape AI trust, though their relative importance may vary across national contexts. To test this framework, we focus on five core mechanisms that previous literature has identified as especially relevant: institutional trust, self-efficacy, perceived threat, technological optimism, and job-related economic vulnerability. These are operationalised as our five main hypotheses below.

H1 (spill-over hypotheses) Trust in AI is not simply a function of attitudes toward the technology itself but also

reflects broader orientations toward the institutional and social environment in which AI operates. Drawing on the notion of trust spill-over, and in-line with trust-transfer perspectives (Goh and Ho 2024; Montag et al. 2023, 2024), we expect that citizens who express greater confidence in institutions and other people will also be more inclined to trust artificial intelligence. We identify three distinct but related sources of such institutional and interpersonal trust:

H1a (government trust effect) Individuals who express higher trust in their national government will report higher trust in AI. This relationship reflects the expectation that government actors can be relied upon to oversee and regulate emerging technologies in the public interest. Prior work suggests that trust in public authorities shapes perceptions of technological legitimacy, particularly where state involvement is visible or expected.

H1b (scientific trust effect) Individuals who express higher trust in university-based scientists will report higher trust in AI. Since many AI developments are either pioneered within, or legitimised by, academic and scientific institutions, we expect that confidence in the scientific community will contribute to higher trust in AI systems. This echoes previous findings that associate trust in science with openness to innovation and acceptance of evidence-based technologies (e.g., Cheung and Ho 2025a).

H1c (social trust effect) Individuals who express higher generalised social trust (namely, the belief that most people can be trusted) will report higher trust in AI. General social trust has long been linked to cooperative norms and a baseline willingness to engage with unfamiliar systems. We argue that this interpersonal trust can also shape attitudes toward abstract systems like AI, particularly in contexts of informational asymmetry where users must take on trust what they cannot fully verify.

Together, these sub-hypotheses articulate a broader claim that trust in AI is shaped by a climate of institutional and interpersonal confidence, rather than being formed in isolation. As such, we expect to find significant positive relationships between each form of spill-over trust and AI trust in both national contexts.

H2 (self-efficacy hypothesis). Individuals with higher AI self-efficacy will express greater trust in AI.

Perceptions of personal competence in understanding and managing AI, commonly described as AI self-efficacy, are increasingly recognised as a significant factor in shaping public trust. Rooted in broader theories of self-efficacy, this construct refers to individuals' belief in their ability to understand, evaluate, and interact with AI systems, and is strongly associated with openness to adoption and confidence in technological systems (see also Bandura 1982; Schwarz and Unsel 2024). Empirical evidence suggests that individuals with higher AI self-efficacy tend to express greater trust in AI. Montag et al. (2023) demonstrate that

individuals with stronger technology self-efficacy are more likely to accept AI technologies and less likely to report fear or anxiety. Importantly, their findings show that this effect is mediated by a broader propensity to trust automation. In other words, the belief that one can competently engage with AI supports both reduced apprehension and increased willingness to delegate decisions to such systems. This is also consistent with the work of Shin (2020, 2021) who shows that trust in AI is shaped by perceptions of fairness, accountability, transparency, and explainability, all of which are closely related to users' sense of understanding and control.

Further support for our argument comes from studies indicating that self-reported knowledge and familiarity with AI are positively associated with trust. For instance, Robles and Mallinson (2025) report that across experimental vignettes, citizens who felt more informed about AI expressed greater support for its use in the scenarios presented, particularly when it was presented in familiar or local institutional contexts. Similarly, Cheung and Ho (2025a) highlight that individuals' perceptions of AI explainability, closely related to the idea of understanding how AI systems work, significantly affect trust across several dimensions, including perceived performance, purpose, and process. While explainability itself is often treated as a system feature, their findings suggest that public perceptions of understandability play an equally important role in fostering trust. Lower AI self-efficacy may not only reduce trust but also be associated with political and epistemic detachment, fueling wider scepticism towards digital governance initiatives (Splichal 2022). This view aligns with findings from risk communication research showing that efficacy beliefs moderate perceived technological threat and shape collective responses to emerging risks (Schwarz and Unsel 2024).

H3 (threat hypothesis). Greater perceived societal threat from AI will be associated with lower trust in AI.

Concerns about the societal risks posed by artificial intelligence are central to public ambivalence and often act as a significant barrier to trust. While AI is frequently promoted as a tool for efficiency, innovation, and problem-solving, citizens increasingly perceive it as a potential source of harm, in social, economic, and ethical terms. These perceived threats range from fears about surveillance and discrimination to anxieties over job loss, algorithmic bias, and erosion of human decision-making authority. Robles and Mallinson (2025) highlight that public trust in AI is not simply a function of technological performance or system design but is deeply shaped by the broader policy and risk environment in which AI is deployed. Similarly, Cheung and Ho (2025b) demonstrate that public trust in AI systems, such as autonomous passenger drones, is significantly undermined when these technologies are perceived as opaque or threatening. Their survey-based study shows that perceptions of uncertainty and danger are associated with heightened concerns

over malfunction, accountability, and misuse, reflecting underlying risk perceptions. Kreps et al. (2023) further underscore the “trust paradox”: individuals may be willing to use AI but do so without corresponding levels of trust, due to perceived risks such as inadequate oversight, system error or displacement of human judgment. These patterns echo broader findings in the risk literature that perceived risk and benefit are inversely related (Siegrist 2000, 2021), implying that AI trust is fundamentally embedded in wider social evaluations of risk, benefit, and governance. These concerns intersect with broader anxieties around power, transparency, and legitimacy, particularly when AI appears imposed without public scrutiny or control (Splichal 2022).

H4 (optimism hypothesis). A more optimistic view of AI’s future benefits will be associated with higher trust in AI.

While concerns about AI are widespread, public trust is also shaped by positive expectations about the technology’s potential to generate societal value. This form of technological optimism is not merely the inverse of fear but a distinct cognitive orientation: a belief that AI will produce meaningful improvements in areas such as healthcare, education, and governance. Robles and Mallinson (2025) find that optimism about AI’s societal benefits, such as reducing administrative burdens or improving public service delivery, can counterbalance concerns about its complexity or ethical ambiguity. Kreps et al. (2023) similarly show that individuals often support AI use not because they fully trust current systems, but because they expect future versions to be more capable and trustworthy. Bao et al. (2022) identify a distinct public segment characterised by strongly optimistic views about AI’s role in solving structural problems, and Cheung and Ho (2025b) show that more positive perceptions of AI’s societal role (for example as a solution to congestion or climate challenges) are associated with higher levels of trust, even amid technical concerns. Optimism thus acts as a kind of benefit-of-the-doubt reasoning, enabling people to tolerate current shortcomings in anticipation of longer-term gains. This reflects what Slovic (1987) termed the “affect heuristic,” where positive emotions about perceived benefits systematically increase trust and reduce perceived risks.

H5 (job-risk hypothesis). Individuals who believe AI will replace their own job will be associated with lower trust in AI.

Concerns about job displacement remain one of the most persistent and personal sources of unease surrounding AI. Even when individuals recognise the broader societal benefits of AI, those who feel personally threatened by it, particularly in terms of economic livelihood, often report significantly lower trust. Ivan (2024), drawing on Eurobarometer data, finds that perceived vulnerability to AI-driven job loss is associated with more pessimistic views of the technology. Bao et al. (2022) likewise identify a population

cluster characterised by job anxiety and economic insecurity, who consistently express lower trust in AI. Robles and Mallinson (2025) report similar findings, noting that even white-collar workers fear the erosion of skill value, autonomy, or decision-making roles. These perceptions also resonate with broader narratives of fairness and economic justice: if AI is viewed as primarily serving employers, governments, or corporations, rather than workers or the public, it is likely to be met with scepticism. As Splichal (2022) warns, when data-driven systems are perceived as tools for optimisation rather than empowerment, public mistrust may extend not only to AI but also to the institutions responsible for its deployment. This link between perceived distributive injustice and institutional distrust is well established in the political economy literature (Bouckaert and van de Walle 2003) underscoring how economic threat perceptions translate into diminished technological trust.

4 Data and methodology

Our data were collected via online surveys in the United Kingdom and Japan. In the United Kingdom, this was done via the YouGov online panel in November 2024 and in the same month in Japan through the Rakuten Insight panel. Full replication data and code are available from the Harvard Dataverse, at <https://doi.org/10.7910/DVN/75AK75>. Our total sample size for the United Kingdom is 1040 respondents and for Japan 2195 respondents, although not all respondents completed all questions, which leads to a lower N for certain parts of the analysis. Table 1 presents descriptive statistics for all variables used in the analyses, together with Welch *t* tests (for the numeric variables) and χ^2 tests (for the binary variables) comparing the Japanese and UK samples. The dependent variable, trust in AI, is measured on a 1–7 scale and was framed as: “Using a scale of 1 to 7 where 1 means ‘Not at all’ and 7 means ‘Completely’, how much do you trust artificial intelligence?” On average, Japanese respondents expressed markedly higher trust than those in the UK. This difference, statistically significant in the Welch test, underlines our interest in cross-national contrasts.

Several key individual-level predictors are included on both theoretical and empirical grounds. Age, gender, and university education serve as baseline demographic controls, capturing socio-structural variation in AI perceptions. The left–right ideology scale (0–10) measures political orientation, which prior research links to technology attitudes, especially in polarised or populist contexts (Araujo et al. 2023; König 2023). Household finances (1–5) provide a proxy for subjective economic wellbeing, a useful indicator given widespread concern that automation may widen economic insecurity.

Table 1 Descriptive statistics

Description	Japan	UK	<i>p</i> (JP vs UK)	Sig
<i>N</i>	2195	1040		
Trust in AI (1–7, mean, SD)	3.41 (1.24)	2.48 (1.33)	< 0.001	***
Age (mean, SD)	49.7 (16.3)	49.8 (16.7)	0.925	
Age (min, max)	18–79	18–90		
Women (%)	48.4%	51.5%	0.099	
University degree (%)	52.4%	36.6%	< 0.001	***
Left–right ideology (0–10, mean, SD)	5.10 (1.56)	4.85 (2.12)	< 0.001	***
Household finances (1–5, mean, SD)	2.83 (0.84)	3.03 (0.90)	< 0.001	***
General social trust (1–7, mean, SD)	3.12 (1.53)	3.52 (1.57)	< 0.001	***
Trust in government (1–7, mean, SD)	3.36 (1.40)	2.85 (1.58)	< 0.001	***
Trust in university scientists (1–7, mean, SD)	4.19 (1.24)	4.76 (1.41)	< 0.001	***
Risk-taking tendency (0–10, mean, SD)	4.56 (1.98)	4.64 (2.18)	0.351	
AI self-efficacy (1–7, mean, SD)	3.79 (1.33)	3.53 (1.74)	< 0.001	***
AI optimism (1–7, mean, SD)	4.30 (1.02)	3.65 (1.41)	< 0.001	***
AI scares me (1–7, mean, SD)	4.48 (1.40)	4.88 (1.58)	< 0.001	***
AI will replace my job (1–7, mean, SD)	3.63 (1.55)	3.24 (1.90)	< 0.001	***

We also incorporate three trust variables (general social trust, trust in government and trust in university scientists) that speak directly to our theoretical claim that trust in AI is associated with broader institutional confidence. The general social trust item (“Generally speaking, would you say that most people can be trusted, or that you can’t be too careful in dealing with people?”) captures baseline interpersonal trust, while trust in university scientists parallels the AI trust question, permitting a direct comparison of technological versus scientific trust.

Beyond these controls, several attitudinal and psychological measures tap cognitive and affective orientations toward AI. AI self-efficacy (agreement with “I know how AI technology can help me”) derives from Bandura’s (1982) self-efficacy framework within social cognitive theory and reflects individuals’ perceived competence in navigating AI systems. As hypothesised, higher efficacy is expected to be associated with higher trust, particularly where AI appears opaque or unfamiliar. Likewise, risk-taking tendency (0–10) gauges openness to uncertainty, a trait often associated with willingness to adopt novel technologies. AI optimism was measured by averaging eight items¹ rated on a 1–7 agreement scale. The items capture enthusiasm,

perceived benefits, and comfort with AI in daily and occupational contexts, and the scale demonstrated excellent internal consistency (Japan $\alpha = 0.91$; UK $\alpha = 0.94$). In contrast, AI fear (“AI scares me”) reflects affective threat perceptions that may dampen trust even where practical benefits are acknowledged. Finally, job replacement anxiety (agreement with “AI will replace my job role entirely”) addresses one of the most personal dimensions of automation anxiety: the perceived threat of displacement.

The cross-tabulated results highlight several statistically significant national contrasts (all at $p < 0.001$). Japanese respondents report greater AI optimism ($M = 4.30$ vs 3.65) and slightly higher AI self-efficacy ($M = 3.79$ vs 3.53), while UK respondents score higher on general social trust ($M = 3.52$ vs 3.12) and trust in university scientists ($M = 4.76$ vs 4.19). Household finances are perceived as stronger in the UK ($M = 3.03$ vs 2.83), and a larger share of Japanese respondents hold university degrees (52.4% vs 36.6%; $p < 0.001$). Fear of AI is more pronounced in the UK ($M = 4.88$ vs 4.48), whereas concern that AI will replace one’s job is higher in Japan ($M = 3.63$ vs 3.24). Risk-taking tendencies and age distributions do not differ significantly across samples.

Overall, these descriptive contrasts indicate that the two publics encounter AI from distinct attitudinal baselines: Japanese respondents combine comparatively high optimism and efficacy with lower institutional trust, while UK respondents exhibit stronger institutional confidence but greater apprehension about AI’s social consequences. Such differences underscore why the subsequent regression analyses are estimated separately by country, allowing us to examine whether the mechanisms linking risk, efficacy, and

¹ I am interested in using artificial intelligence systems in my daily life; there are many beneficial applications of artificial intelligence; artificial intelligence is exciting; artificial intelligence can provide new economic opportunities for this country; I would like to use artificial intelligence in my job; an artificial intelligence agent would be better than an employee in many routine jobs; I am impressed by what artificial intelligence can do; I am comfortable working alongside AI (i.e., generative AI, chatbots or robots) in my current and future jobs.

trust operate similarly across divergent cultural and institutional contexts.

5 Analysis

We begin by testing the five core hypotheses outlined above using linear regression models for each country, treating the 1–7 trust scale as approximately continuous. Table 2 presents the results for Japan (Models 1–5) and the UK (Models 6–10). Each column sequentially adds blocks of variables corresponding to our theoretical framework, beginning with institutional and interpersonal trust (H1a–c, Models 1 and 6), followed by self-efficacy (H2, Models 2 and 7), perceived threat (H3, Models 3 and 8), technological optimism (H4, Models 4 and 9), and economic vulnerability (H5, Models 5 and 10). Variance inflation factors (VIFs) were inspected to assess potential multicollinearity. All predictors were well within acceptable limits (Japan: max VIF = 1.42; UK: max VIF = 1.84), indicating no evidence of problematic multicollinearity.

The results provide strong support for the Spill-over Hypotheses (H1a–c). Across all model specifications and both countries, higher trust in government, trust in university scientists, and generalised social trust are consistently and significantly associated with higher levels of trust in AI. This pattern holds even after accounting for a wide range of individual-level controls, suggesting that institutional and interpersonal trust exert an independent and stable influence on AI attitudes. Notably, the strength and direction of these effects are highly similar across both national contexts, reinforcing the notion that trust in AI is embedded within broader social and institutional trust architectures. These findings are also in line with the existing research (e.g., Araujo et al. 2023; Dreksler et al. 2025; Kreps et al. 2023; Novozhilova et al. 2024).

Self-efficacy (H2) also emerges as a significant and positive predictor in both Japan and the UK. Individuals who feel confident in their ability to understand how AI can help them are more likely to express trust in AI systems, supporting the view that perceived competence facilitates acceptance. The strength of this association remains relatively stable in Japan, but becomes less robust in the UK after introducing other predictors, highlighting the importance of self-efficacy in shaping AI-related attitudes, in line with Bandura's (1982) seminal argument about the role of self-efficacy.

The effects of self-efficacy should be considered alongside technological optimism (H4). While AI self-efficacy is positively associated with trust in AI in both countries, its role differs across national contexts. In Japan, self-efficacy remains a significant predictor even after accounting for technological optimism, indicating a combination of direct and indirect pathways. In the UK, by contrast, introducing

technological optimism largely accounts for the association between self-efficacy and trust, rendering the direct effect statistically indistinguishable from zero. Figure 1 presents a mediation analysis following Imai et al. (2011), which confirms a statistically significant indirect association between self-efficacy and trust via technological optimism in both countries, with the indirect pathway accounting for a larger share of the total effect in the UK than in Japan. Together, these results suggest that confidence in one's ability to engage with AI fosters trust primarily by shaping forward-looking expectations about AI's societal benefits, particularly in the UK context (Bao et al. 2022; Cheung and Ho 2025a, b).

Support for the Threat Hypothesis (H3) is evident in both countries. Respondents who perceive AI as threatening are significantly less trusting of it. The effect size is slightly larger in the UK, although the direction is consistent across contexts. This aligns with the literature suggesting that affective risk perceptions are a key barrier to technological trust (Siegrist 2021; Krieger et al. 2024).

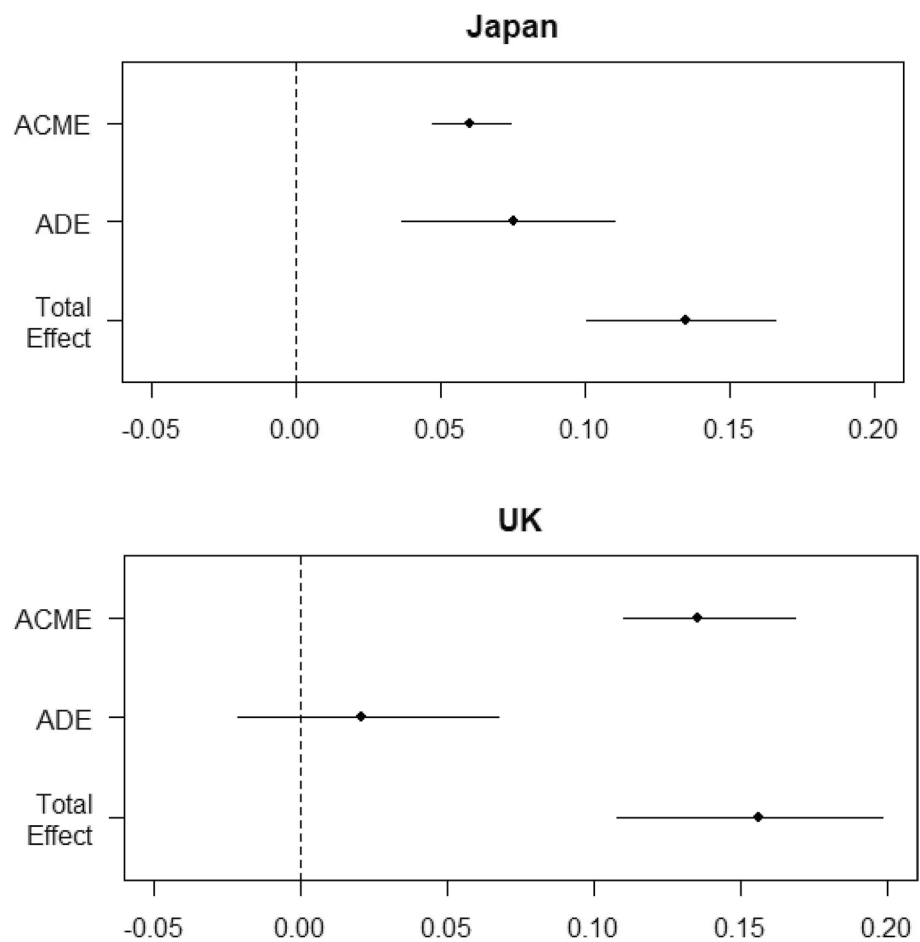
However, the results for the Job Risk Hypothesis (H5) are more mixed, and theoretically revealing. It should be noted that Models 5 and 10 are estimated on a smaller sample, as the job replacement item is only applicable to respondents currently in employment; retired, unemployed, and student respondents are therefore excluded. For this reason, the job-risk variable is introduced in the final model for each country. In Japan, the belief that AI will replace one's job is *positively* associated with trust in AI, even after controlling for all other factors. This association remains positive and statistically significant across alternative specifications. In contrast, in the UK, the results are less consistent, with significance and coefficient direction both unstable (see Appendix). This counterintuitive pattern suggests that perceptions of job replacement reflect multiple, intertwined considerations, including expectations about technological progress, labour market adaptation, and the role of institutions in managing economic change.

To explore this further, we introduced models including an interaction between perceived job replacement risk and risk tolerance to examine changes in marginal effects. One notable finding concerns this interaction. The analysis reveals that, conditional on risk tolerance, the relationship between job replacement perceptions and trust in AI differs across countries among respondents with high-risk tolerance. Among respondents with low-risk tolerance, job replacement perceptions show little association with trust in AI in either country. Among those with high-risk tolerance, however, UK respondents exhibit a pattern consistent with the job-risk hypothesis. In contrast, in Japan, the perception that AI will take one's job remains positively associated with trust in AI even at higher levels of risk tolerance. This divergence highlights the importance of context in shaping

Table 2 Regression models

	Dependent variable: trust in AI									
	Japan					United Kingdom				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Age	0.000 (0.001)	0.003+ (0.001)	0.004** (0.002)	0.004** (0.001)	0.005*** (0.002)	− 0.011*** (0.002)	− 0.003 (0.002)	− 0.003 (0.002)	− 0.003 (0.002)	− 0.003 (0.003)
Women	− 0.007 (0.050)	0.031 (0.049)	0.052 (0.049)	0.054 (0.048)	0.034 (0.051)	− 0.175* (0.077)	− 0.093 (0.074)	0.044 (0.073)	0.083 (0.068)	0.136+ (0.078)
University	− 0.160** (0.050)	− 0.169*** (0.049)	− 0.161** (0.049)	− 0.186*** (0.048)	− 0.188*** (0.051)	− 0.058 (0.085)	− 0.106 (0.082)	− 0.119 (0.078)	− 0.223** (0.074)	− 0.227** (0.083)
Left-right	− 0.004 (0.015)	− 0.008 (0.015)	− 0.004 (0.015)	− 0.008 (0.015)	− 0.008 (0.016)	0.092*** (0.020)	0.080*** (0.019)	0.081*** (0.018)	0.066*** (0.017)	0.065*** (0.019)
Household finance	0.005 (0.030)	− 0.010 (0.029)	− 0.011 (0.029)	− 0.024 (0.029)	− 0.032 (0.031)	0.028 (0.045)	0.006 (0.043)	− 0.019 (0.041)	− 0.059 (0.039)	− 0.007 (0.044)
Trust people	0.110*** (0.017)	0.108*** (0.017)	0.102*** (0.016)	0.111*** (0.016)	0.114*** (0.017)	0.097*** (0.027)	0.099*** (0.026)	0.098*** (0.025)	0.094*** (0.023)	0.088** (0.027)
Trust government	0.149*** (0.019)	0.147*** (0.019)	0.142*** (0.018)	0.140*** (0.018)	0.119*** (0.019)	0.163*** (0.028)	0.136*** (0.027)	0.118*** (0.026)	0.102*** (0.024)	0.121*** (0.028)
Trust university scientists	0.294*** (0.021)	0.278*** (0.021)	0.289*** (0.021)	0.252*** (0.021)	0.254*** (0.022)	0.143*** (0.031)	0.136*** (0.030)	0.131*** (0.028)	0.097*** (0.027)	0.085** (0.032)
Take risks	0.038** (0.012)	0.024* (0.012)	0.021+ (0.012)	0.008 (0.012)	0.009 (0.013)	0.053** (0.018)	0.031+ (0.017)	0.023 (0.017)	0.013 (0.016)	0.032+ (0.019)
AI self-efficacy	0.140*** (0.018)	0.140*** (0.018)	0.131*** (0.018)	0.070*** (0.020)	0.054* (0.021)	0.018 (0.021)	0.232*** (0.023)	0.154*** (0.023)	0.020 (0.025)	0.034 (0.028)
AI scares me			− 0.107*** (0.017)	− 0.097*** (0.017)	− 0.098*** (0.018)			− 0.234*** (0.025)	− 0.160*** (0.024)	− 0.177*** (0.029)
AI optimistic future				0.219*** (0.026)	0.205*** (0.028)				0.370*** (0.031)	0.347*** (0.035)
AI will replace my job					0.072*** (0.016)				0.029 (0.021)	0.029 (0.021)
Constant	1.232*** (0.148)	0.786*** (0.157)	1.197*** (0.169)	0.694*** (0.177)	0.593** (0.195)	0.861*** (0.257)	− 0.032 (0.260)	1.476*** (0.298)	0.736* (0.286)	0.487 (0.339)
Num.Obs	2195	2195	2195	2195	1973	1040	1040	1040	1040	783
R ²	0.213	0.233	0.247	0.271	0.265	0.162	0.239	0.298	0.382	0.395
R ² Adj	0.209	0.229	0.243	0.267	0.260	0.155	0.231	0.290	0.375	0.385
F	65.567	66.311	64.986	67.472	54.208	22.105	32.253	39.624	53.006	38.583

Fig. 1 Mediation analysis of the relationship between AI self-efficacy and trust in AI via technological optimism (Japan and the UK)



how potential economic vulnerability is experienced and interpreted in relation to emerging technologies (see also Novozhilova et al. 2024).

Together, these findings offer broad support for the central claim that trust in AI is structured by both societal-level trust orientations and individual-level psychological dispositions. While many predictors operate similarly across countries, the contrasting patterns in job replacement perceptions underscore the value of cross-national comparison in unpacking the contingencies of public trust.

6 Discussion

This study examined the individual and institutional foundations of public trust in artificial intelligence in Japan and the UK. The results confirm that both psychological predispositions, such as AI self-efficacy and technological optimism, and broader patterns of institutional and interpersonal trust play key roles in shaping public confidence in AI. Yet, the two countries diverge in the consistency and magnitude of some effects, revealing important context-specific dynamics.

In both countries, trust in government, trust in university scientists, and generalised social trust emerged as the most reliable predictors of trust in AI, providing strong and consistent support for the Spill-over Hypotheses (H1a–c). This underscores the idea that AI technologies are not evaluated in isolation but through a broader lens of institutional legitimacy and social trust. Citizens who trust the institutions and social actors surrounding AI (whether political, scientific, or interpersonal) are more inclined to trust the technology itself. The stability of these relationships across countries and model specifications points to a foundational role for trust spill-over in shaping AI attitudes.

The role of AI self-efficacy (H2) also aligns with theoretical expectations: individuals who feel confident in their ability to use and understand AI express more trust in it. This effect was especially stable in Japan, while in the UK it showed a drop-off in significance once other controls were introduced. However, the mediation analysis indicates that the association between self-efficacy and trust operates in part via technological optimism. This interpretation is consistent with research indicating that UK publics exhibit more conditional trust in digital innovation, particularly where personal or ethical concerns are salient (Horvath et al. 2023).

Affective orientations toward AI, particularly optimism and fear, also proved important. Optimism (H4) was a significant and positive predictor of trust in both countries, although its impact was slightly stronger in the UK. As noted earlier, in the UK, technological optimism plays a mediating role in the relationship between self-efficacy and trust in AI. In addition, fear of AI (H3), measured by agreement with the statement “AI scares me”, was more strongly negative in the UK, lending support to the idea that public discourse in Britain has framed AI more as a disruptive or alienating force (House of Lords 2018). These findings align with prior research suggesting that emotional appraisals exert a powerful influence on technological trust, even when controlling for institutional and cognitive factors (Glikson and Woolley 2020; Riley and Dixon 2024).

The clearest cross-national divergence appears in perceptions of job replacement (H5). In Japan, the belief that AI will replace one’s job is positively associated with trust in AI; a counterintuitive but consistent result. This may reflect cultural narratives of resilience, adaptation, or even technological harmony, in a society facing acute demographic ageing and labour shortages. In such a context, automation may be viewed less as a threat than as a pragmatic necessity. In other words, individuals may feel they have little choice but to actively embrace AI as a tool capable of performing work that would otherwise remain undone. In the UK, by contrast, no significant relationship was observed. Among UK respondents with high-risk tolerance, advanced AI may be perceived more as a competitive hindrance than as an opportunity. However, unlike previous studies, the present analysis finds no clear evidence that perceived job displacement leads to distrust in AI. These findings carry several broader implications. First, they suggest that public trust in AI cannot be fostered solely through technical design, transparency or user education. Rather, trust is embedded in a wider constellation of beliefs about institutions, expertise, and social cohesion. Second, national context matters, not only in shaping which factors are most influential, but also in conditioning how citizens interpret them (see also Novozhilova et al. 2024). The contrast between Japan’s relatively stable trust architecture and the UK’s more contingent, affect-driven pattern highlights the need for context-aware governance and communication strategies. Finally, the divergence in perceptions of job displacement points to the importance of labour market narratives and economic imaginaries in shaping technological legitimacy. As societies negotiate the risks and benefits of AI adoption, understanding how publics interpret potential economic vulnerability will be key to maintaining trust.

Before turning to the policy implications of our findings, we also acknowledge several limitations of this research. As argued by Dreksler et al. (2025), trust in AI can be measured in a number of different ways and even what “artificial

intelligence” means for the respondent can be problematic. Our results are based on the broadest possible definition of AI and as such we are not able to disentangle the particular types of AI and the perceptions thereof. This is certainly a limitation, although it should be noted that the public’s understanding of different types of AI is unlikely to be detailed enough to allow for precise judgements of each. In that sense, using a broader conceptual framing, which also reflects how AI is typically discussed in the media, still provides meaningful insight. Another limitation is that we use attitudinal measures without an experimental setup. This means that we cannot claim causal relationships, but instead identify associations between the variables of interest. Some research has drawn on experiments (e.g., Kreps et al. 2023) to ascertain the causal mechanism in place related to trust in AI; for future work, this strand of research can be developed further.

These findings suggest that trust in AI reflects not only individual attitudes but also the broader legitimacy of governance systems. While public confidence in AI can facilitate responsible adoption, trust is not an unqualified good: where oversight is weak or risks are poorly communicated, misplaced trust may amplify harm (Floridi 2023; Hagendorff 2023). Policies that strengthen transparency, accountability, and explainability (principles central to the EU AI Act and related frameworks) are therefore essential to maintaining trust only where it is deserved, and withholding it where it is not. Public engagement and education initiatives that enhance AI self-efficacy may also foster informed optimism, helping citizens evaluate potential benefits and risks more critically. Finally, the complex pattern observed for job-risk perceptions underscores that economic vulnerability and technological optimism can coexist, reflecting both anxiety and adaptation. Future governance and communication strategies should recognise this duality: public trust in AI will be most resilient when it is built not on blind confidence, but on a well-informed sense of capability and control.

6.1 Policy implications

Despite the limitations related to our findings, they nevertheless offer several practical implications for policymakers, particularly those concerned with designing and regulating AI systems in ways that secure public legitimacy. First, the consistent effects of institutional and interpersonal trust (specifically, trust in government, university scientists, and even other people) highlight the importance of embedding AI initiatives within a broader climate of social and institutional credibility. Where confidence in public institutions and scientific actors is low, or where social cohesion is weak, trust in AI is unlikely to flourish, even if the technology itself is well designed. This underscores the need for joined-up

governance, where AI policy is framed not as a technical fix but as part of a trustworthy institutional ecosystem.

Second, the strong association between trust and AI self-efficacy, especially in Japan, points to a clear target for public engagement. Boosting people's confidence in their ability to understand and interact with AI can help to demystify the technology and increase support for its adoption. Public education campaigns, workplace training and civic participation in AI development and oversight may help foster a sense of ownership and capability. While self-efficacy proved less robust in the UK once other factors were controlled, this may reflect a need for such interventions to address contextual fears and anxieties, rather than focusing solely on technical understanding (Shin 2021).

Third, the findings emphasise the need for tailored risk communication. While affective orientations like fear had negative effects in both countries, they were especially pronounced in the UK. This suggests that UK policymakers and civil society actors should prioritise messaging strategies that emphasise explainability, transparency, and ethical safeguards, particularly in domains such as employment, healthcare, or surveillance, where AI is often perceived as intrusive. At the same time, framing AI as a contributor to shared societal goals, from climate adaptation to health system resilience, may temper anxiety and build constructive engagement (see also Siegrist 2021; Krieger et al. 2024).

Finally, the divergent role of employment-related concerns suggests that AI governance must be sensitive to national labour market narratives. In Japan, where the belief that AI would replace one's own job was positively associated with trust, public discourse may already frame automation as a pragmatic response to demographic pressures and labour shortages. In contrast, no such relationship emerged in the UK once controls were applied, pointing to deeper ambivalence about technological change. Here, trust may depend less on automation's utility and more on whether citizens feel protected against its consequences. Policymakers should therefore pair innovation strategies with robust safety nets, upskilling opportunities, and inclusive narratives about economic transition, ensuring that public trust in AI is not undermined by fears of personal or collective precarity. At the same time, policymakers should recognise that forward-looking optimism about the future remains an important foundation for building trust in AI.

7 Conclusion

This study has examined the psychological, institutional, and affective underpinnings of public trust in artificial intelligence across two national contexts: Japan and the United Kingdom. Drawing on survey data from over 3000 respondents, we tested a series of hypotheses linking AI trust to

broader patterns of social and political trust, individual self-efficacy, optimism, fear, and perceived job threat.

Our findings support a multi-dimensional account of AI trust, in which attitudes toward AI are not formed in isolation but shaped by a broader ecosystem of beliefs about institutions, expertise, personal competence, and risk. The most consistent predictors, trust in government, trust in university scientists, and generalised social trust, underscore the importance of a spill-over logic, in which institutional and interpersonal confidence extends into the technological domain. This suggests that building public trust in AI will depend not only on how AI systems function, but on who is seen to design, deploy and regulate them. We also find strong evidence for the role of self-efficacy and affective orientation. Individuals who feel capable of understanding and using AI, and those who express optimism about its societal role, are significantly more likely to trust it. Fear, by contrast, erodes trust, particularly in the UK, where public discourse around AI is more conflictual and risk-oriented. These dynamics reveal that trust is not just cognitive or institutional but also deeply emotional, shaped by feelings of vulnerability, hope, and control.

Finally, our results point to a clear cross-national difference: in Japan, believing that AI will replace one's job is positively associated with trust in AI, suggesting a narrative of automation as beneficial or even necessary. In the UK, no such relationship emerges. This divergence highlights the importance of national context and cultural framing in shaping public attitudes toward technology. Where AI is embedded in narratives of national renewal or problem-solving, trust may follow, even in the face of economic disruption. Overall, our study contributes to a growing literature on the social foundations of AI legitimacy. It reinforces the view that AI governance is not just a technical or regulatory task but a deeply political one, requiring careful attention to the public meanings of technology, trust, and institutional authority. As AI becomes increasingly embedded in public life, understanding and engaging with these meanings will be critical to ensuring its democratic acceptance and responsible use.

Acknowledgements We would like to thank Han Dorussen, Naofumi Fujimura, Naoko Matsumura, Jason Reifler, Thomas Scotto, Atsushi Tago, Dorothy Yen and Masahiro Zenkyo for their help in grant acquisition and survey development.

Author contributions All authors engaged in survey design. SDP and YS managed the survey and the project. All authors wrote, revised and reviewed the manuscript.

Funding This article was funded by Japan Society for the Promotion of Science, JPJSJRP 20211704, Economic and Social Research Council, ES/W011913/1.

Data availability Full replication data and code are available from the Harvard Dataverse, at: <https://doi.org/10.7910/DVN/75AK75>.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Araujo T, Brosius A, Goldberg AC, Möller J, de Vreese C (2023) Humans vs. AI: the role of trust, political attitudes, and individual characteristics on perceptions about automated decision making across Europe. *Int J Commun* 17:28
- Bandura A (1982) Self-efficacy mechanism in human agency. *Am Psychol* 37(2):122
- Bao L, Krause NM, Calice MN, Scheufele DA, Wirz CD, Brossard D et al (2022) Whose AI? How different publics think about AI and its social impacts. *Comput Hum Behav* 130:107182
- Bouckaert G, Van de Walle S (2003) Comparing measures of citizen trust and user satisfaction as indicators of “good governance”: difficulties in linking trust and satisfaction indicators. *Int Rev Adm Sci* 69(3):329–343
- Cheung JC, Ho SS (2025a) The effectiveness of explainable AI on human factors in trust models. *Sci Rep* 15(1):23337
- Cheung JC, Ho SS (2025b) Explainable AI and trust: how news media shapes public support for AI-powered autonomous passenger drones. *Public Underst Sci* 34(3):344–362
- Curtice J, Montagu I (2022) Is trust in government declining? Evidence from Britain. *British Social Attitudes* 39, National Centre for Social Research. <https://natcen.ac.uk>. Accessed 26 Aug 2026.
- David P, Choung H, Seberger JS (2024) Who is responsible? US public perceptions of AI governance through the lenses of trust and ethics. *Public Underst Sci* 33(5):654–672
- Domagala N (2021) What is our new algorithmic transparency standard? Data in government blog. <https://dataingovernment.blog.gov.uk/2021/11/29/what-is-our-new-algorithmic-transparency-standard/>. Accessed 26 Aug 2026.
- Dreksler N, Law H, Ahn C, Schiff D, Schiff KJ, Peskowitz Z (2025) What does the public think about AI? An overview of the public's attitudes towards AI and a resource for future research. <https://ssrn.com/abstract=5108572> or <https://doi.org/10.2139/ssrn.5108572>. Accessed 26 Aug 2026.
- Floridi L (2023) The ethics of artificial intelligence: principles, challenges, and opportunities. Oxford University Press, Oxford
- Gillespie N, Lockey S, Ward T, Macdade A, Hassed G (2025) Trust, attitudes and use of artificial intelligence: a global study 2025. Univ Melb KPMG. <https://doi.org/10.26188/28822919>
- Glikson E, Woolley AW (2020) Human trust in artificial intelligence: review of empirical research. *Acad Manag Ann* 14(2):627–660
- Goh TJ, Ho SS (2024) Trustworthiness of policymakers, technology developers, and media organizations involved in introducing AI for autonomous vehicles: a public perspective. *Sci Commun*. <https://doi.org/10.1177/10755470241248169>
- Hagendorff T (2023) AI ethics and its pitfalls: not living up to its own standards? *AI Ethics* 3(1):329–336
- Horvath L, James O, Banducci S, Beduschi A (2023) Citizens' acceptance of artificial intelligence in public services: evidence from a conjoint experiment about processing permit applications. *Gov Inf Q* 40(4):101876
- House of Lords Select Committee on Artificial Intelligence (2018) AI in the UK: ready, willing and able? <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf>. Accessed 26 Aug 2026.
- Imai K, Keel L, Tingley D, Yamamoto T (2011) Unpacking the black box of causality: learning about causal mechanisms from experimental and observational studies. *Am Polit Sci Rev* 105(4):765–789
- Ivan B (2024) European public attitudes and perceptions on digitalization and artificial intelligence. *J Compar Res Anthropol Sociol* 15(1):57–79
- König PD (2023) Citizen conceptions of democracy and support for artificial intelligence in government and politics. *Eur J Polit Res* 62(4):1280–1300
- Kreps S, George J, Lushenko P, Rao A (2023) Exploring the artificial intelligence “trust paradox”: evidence from a survey experiment in the United States. *PLoS ONE* 18(7):e0288109
- Krieger JB, Boudier F, Wibral M, Almeida RJ (2024) A systematic literature review on risk perception of artificial narrow intelligence. *J Risk Res* 28:1–19
- Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. *Hum Fact* 46(1):50–80
- Levi M, Stoker L (2000) Political trust and trustworthiness. *Annu Rev Polit Sci* 3:475–507
- Liehner GL, Biermann H, Hick A, Brauner P, Zieffle M (2023) Perceptions, attitudes and trust towards artificial intelligence—an assessment of the public opinion. *Artif Intell Soc Comput* 72:32–41
- Mayer RC, Davis JH, Schoorman FD (1995) An integrative model of organizational trust. *Acad Manag rev* 20(3):709–734.
- METI (Ministry of Economy, Trade and Industry of Japan) (2021) AI governance guidelines. <https://www.meti.go.jp>. Accessed 26 Aug 2026.
- Montag C, Kraus J, Baumann M, Rozgonjuk D (2023) 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
- Montag C, Becker B, Li BJ (2024) 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(2):100070
- Neri H, Cozman F (2020) The role of experts in the public perception of risk of artificial intelligence. *AI Soc* 35(3):663–673
- Newton K, Zmerli S (2011) Three forms of trust and their association. *Eur Polit Sci Rev* 3(2):169–200
- Nguyen D (2023) How news media frame data risks in their coverage of big data and AI. *Internet Policy Rev* 12(2):1–30
- Nguyen D, Hekman E (2024) The news framing of artificial intelligence: a critical exploration of how media discourses make sense of automation. *AI Soc* 39(2):437–451
- Norris P (2011) Democratic deficit: critical citizens revisited. Cambridge University Press, Cambridge
- Novozhilova E, Mays K, Paik S, Katz JE (2024) More capable, less benevolent: trust perceptions of AI systems across societal contexts. *Mach Learn Knowl Extract* 6(1):342–366
- OECD (2021) Government at a glance 2021. OECD Publishing, Paris. <https://doi.org/10.1787/1c258f55-en>
- OECD (2023) Trust in government: evidence across OECD countries. OECD Publishing, Paris

- Riley BK, Dixon A (2024) Emotional and cognitive trust in artificial intelligence: a framework for identifying research opportunities. *Curr Opin Psychol*. <https://doi.org/10.1016/j.copsyc.2024.101833>
- Robles P, Mallinson DJ (2025) Artificial intelligence technology, public trust, and effective governance. *Rev Policy Res* 42(1):11–28
- Robles P, Mallinson DJ (2023) Catching up with AI: Pushing toward a cohesive governance framework. *Politics & Policy*. 51(3):355–372.
- Rojahn J, Palu A, Skiena S, Jones JJ (2023) American public opinion on artificial intelligence in healthcare. *PLoS ONE* 18(11):e0294028
- Schiff DS (2024) Framing contestation and public influence on policy-makers: evidence from US artificial intelligence policy discourse. *Policy Soc* 43(3):255–288
- Schiff KJ, Schiff DS, Adams IT, McCrain J, Mourtgos SM (2025) Institutional factors driving citizen perceptions of AI in government: evidence from a survey experiment on policing. *Public Adm Rev* 85(2):451–467
- Schwarz A, Unselt JJ (2024) Rage against the machine? Framing societal threat and efficacy in YouTube videos about artificial intelligence. *Risk Anal* 44(10):2377–2395
- Schwarz A (2024) The mediated amplification of societal risk and risk governance of artificial intelligence: technological risk frames on YouTube and their impact before and after ChatGPT. *J Risk Res*:1–23
- Shin D (2020) User perceptions of algorithmic decisions in the personalized AI system: perceptual evaluation of fairness, accountability, transparency, and explainability. *J Broadcast Electron Media* 64(4):541–565
- Shin D (2021) The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. *Int J Hum-Comput Stud* 146:102551
- Siau K, Wang W (2018) Building trust in artificial intelligence, machine learning, and robotics. *Cutter Bus Technol J* 31(2):47–53
- Siegrist M (2000) The influence of trust and perceptions of risks and benefits on the acceptance of gene technology. *Risk anal* 20(2):195–204.
- Siegrist M (2021) Trust and risk perception: A critical review of the literature. *Risk anal* 41(3):480–490.
- Slovic P (1987) Perception of risk. *Science*. 236(4799):280–285.
- Splichal S (2022) In data we (don't) trust: the public adrift in data-driven public opinion models. *Big Data Soc* 9(1):20539517221097320
- Thiebes S, Lins S, Sunyaev A (2021) Trustworthy artificial intelligence. *Electron Markets* 31:447–464. <https://doi.org/10.1007/s12525-020-00441-4>
- Van der Meer TWG, Ouattara E (2019) Putting 'political' back in political trust: an IRT test of the unidimensionality and cross-national equivalence of political trust measures. *Qual Quant* 53(6):2983–3002
- WEF/Ipsos (2024) Global views on artificial intelligence 2024. World Economic Forum and Ipsos. <https://www.weforum.org>. Accessed 26 Aug 2026.
- Zmerli S, Newton K (2008) Social trust and attitudes toward democracy. *Public Opin Q* 72:706–724

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