ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute
Discussion Paper No. 417

Human Trust in AI: Evidence from Experimental Economics

Bernd Irlenbusch

June 2026

www.econtribute.de

UNIVERSITÄT BONN

UNIVERSITÄT ZU KÖLN

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/2 – 390838866 – is gratefully acknowledged.

Human Trust in AI: Evidence from Experimental Economics

Bernd Irlenbusch*

June 2026

Abstract

Artificial intelligence increasingly shapes economic decisions, yet its value depends on whether humans rely on it appropriately. This survey selectively reviews experimental economic evidence (2020 – 2026) on trust in AI, with a focus on privacy, transparency, accountability, fairness, and efficiency. The evidence challenges simple accounts of algorithm aversion or algorithm appreciation. Individuals may underuse beneficial AI because of opacity, autonomy concerns, or institutional distrust, but may also over-rely on deficient systems, disclose excessive data, or delegate responsibility strategically. The survey suggests that trust in AI is best understood as calibrated reliance under informational and institutional constraints. Effective governance should structure informational and institutional environments that help humans calibrate reliance on AI to its actual capabilities, limitations, and social consequences.

JEL classification: C90, C91, C92, C93, O33

Keywords: trust in AI; calibrated reliance; algorithm aversion; algorithm appreciation; privacy; transparency; accountability; fairness; efficiency

* Center for Social and Economic Behavior and Department of Corporate Development and Business Ethics at the University of Cologne, Albertus-Magnus-Platz, 50923 Köln, Germany; ORCID: 0000-0002-3433-2826. I thank Carlotta Bach for outstanding student research assistance. I used Google NotebookLM (notebooklm.google.com) to search for and summarize relevant literature, ChatGPT (chatgpt.com) to improve the style and readability of the text, and refine (www.refine.ink) to receive critical feedback on an earlier version of the paper. This survey was primarily written during my sabbatical in the winter semester 2025/2026, which I spent at the Department of Management at the London School of Economics and Political Science – I am enormously grateful for its warm hospitality. Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC2126/2 – 390838866 and the Center for Social and Economic Behavior at the University of Cologne is gratefully acknowledged.

Introduction

Artificial intelligence is increasingly used to allocate opportunities, screen applicants, provide advice, personalize services, support professional judgment, and mediate social and strategic interaction (Capraro et al., 2024). In each of these settings, the economic value of AI depends not only on technical performance but also on whether humans rely on, contest, delegate to, or disclose information to algorithmic systems. Trust is not a peripheral psychological attitude toward technology. It is a behavioral mechanism that shapes disclosure, participation, delegation, contestation, organizational performance, and the distribution of gains from automation.

Trust is commonly defined as a willingness to accept vulnerability based on positive expectations regarding another actor's behavior (Rousseau et al., 1998). In Mayer, Davis, and Schoorman's (1995) seminal model, trust involves willingness to be vulnerable to another party's actions under limited monitoring or control. Trustworthiness, by contrast, refers to properties of the trustee, such as ability, benevolence, and integrity. Duran and Pozzi (2025) further distinguish trust from reliance: reliance may be based on observed regularities, while trust often involves expectations of benevolence, mutual obligations, and shared norms (for further discussions, see Ryan, 2020, and Hahne and Schmoelz, 2026). This distinction is important conceptually and normatively, particularly in philosophical debates about whether AI systems can properly be objects of trust. For the purposes of this survey, however, "trust in AI" is used in a behavioral-economic sense to refer to reliance on AI systems under vulnerability, uncertainty, and limited control, while assessing whether such reliance is warranted by system performance, institutional safeguards, and social consequences.

This survey starts from a simple premise: trust in AI should not be understood as either generalized algorithm aversion or generalized algorithm appreciation. Humans do not uniformly distrust algorithms, nor do they uniformly defer to them. Instead, they rely on AI selectively and contextually. Research has documented algorithm aversion, in which users avoid algorithms after seeing them err (Dietvorst et al., 2015), and algorithm appreciation, in which users prefer algorithmic over human judgment in some settings (Logg et al., 2019). Integrative work emphasizes that these are not opposing traits but context-dependent patterns of reliance (Jussupow, Benbasat, and Heinzl, 2024). The central question is when reliance is warranted, when it is misplaced, and how institutions shape this calibration. Jacovi et al. (2021) formalize this as warranted or "contractual" trust: trust should be causally connected to actual system trustworthiness if misuse and disuse are to be avoided. Trust in AI is best understood as calibrated reliance under informational and institutional constraints.

Calibration means relying on AI when it improves decisions and qualifying or withholding reliance when outputs are unreliable, biased, manipulative, or insufficiently accountable. It cannot be inferred from use,

acceptance, disclosure, or delegation alone; behavior must be assessed against evidence of system trustworthiness while allowing for independent values and institutional constraints. Under-reliance occurs when useful AI is rejected despite credible evidence of performance. Over-reliance occurs when people follow, disclose to, or delegate to AI despite evidence of unreliability, bias, excessive data risk, or responsibility evasion. Resistance to AI should not always be interpreted as irrational aversion: it may express principled resistance grounded in agency, dignity, or deontological concerns (Hannon, Ciriello, and Gal, 2024). Conversely, design efforts should avoid fostering excessive reliance; responsible AI design should aim at appropriate rather than maximal trust (Liao and Sundar, 2022; Mehrotra, Degachi, and Vereschak, 2024). Sociotechnical contexts and individual differences also shape trust formation, including the extent to which users prioritize transparency, capability, and institutional safeguards, as well as how AI understanding, personality, and general propensity to trust affect reliance (Dang & Li, 2026; Orbán & Stefkovics, 2025; Riedl, 2022).

Experimental economics is well-suited to studying these mechanisms because it can separate beliefs from preferences, vary incentives and information structures, and identify causal effects of disclosure, explanation, feedback, delegation rights, and accountability rules. This matters because AI often enters economic life through environments in which people must decide whether, how, and when to rely on it. Individuals respond strategically to algorithms, infer motives from institutional design, update willingness to share data or accept recommendations, and decide whether to retain or relinquish control. Experimental economics can study trust in AI not only as an attitude, but as behavior under incentives, uncertainty, and institutional constraints.

This survey selectively reviews experimental economic evidence on human trust in AI, mainly from 2020 to 2026. It contributes to existing interdisciplinary reviews of AI trust, algorithm aversion, algorithm appreciation, decision automation, and human-machine interaction (Aquilino et al., 2025; Bacha et al., 2024; Burton, Stein, and Jensen, 2020; Chugunova and Sele, 2022; Glikson and Woolley, 2020; Jussupow, Benbasat, and Heinzl, 2020; Kaplan et al., 2023; Langer and Landers, 2021; Mahmud et al., 2022), as well as bibliometric, management, taxonomic, and cognitive accounts of trust in AI (Benk et al., 2025; Henrique and Santos, 2024; Caro et al., 2026; Alekseev and Strobel, 2026; Jiang, Li, and Liu, 2025).

The survey contributes in three ways. First, it reframes trust in AI as calibrated reliance rather than as acceptance, adoption, or favorable attitude. Second, it organizes experimental economic evidence across five domains—privacy, transparency, accountability, fairness, and efficiency—that jointly determine whether reliance is warranted. Third, it derives governance implications for designing informational and institutional environments that make reliance more warranted rather than merely more frequent. This perspective is especially relevant in the post-ChatGPT era, where AI systems increasingly mediate communication, hiring, education, workplace interaction, and professional advice, and where phenomena such as AI-mediated

interaction, generative-model self-preference (Lehr, Cipperman, and Banaji, 2026), AI overconfidence (Sun et al., 2025), and AI sycophancy (Cheng et al., 2026) raise new questions for calibrated reliance.

The survey's core evidence comes from incentivized laboratory and online experiments as well as field experiments. Because trust is rarely observed directly, the reviewed studies typically rely on behavioral proxies such as disclosure, acceptance, advice-taking, delegation, participation, willingness to pay, or compliance. These behaviors are informative only when interpreted relative to the incentives, information, and institutional context of the experiment. The studies deal with economically consequential decisions involving algorithmic or AI-mediated systems, including data sharing, participation in AI-mediated processes, acceptance of recommendations, delegation of authority, allocation, performance, and welfare. Studies from management, information systems, human-computer interaction, and related fields are included when they meet standards associated with experimental economics, such as incentivization and no deception, and identify how incentives, beliefs, information, institutions, or strategic environments shape behavior.

Table 1. Five domains of calibrated reliance in AI

Domain	Behavioral margin	Calibration problem	Governance implication
Privacy	Disclosure and data sharing	Over-disclosure, defensive non-disclosure, or disclosure under strategic pressure	Make data consequences intelligible and govern data externalities credibly
Transparency	Advice-taking, acceptance, and contestation	Source aversion, anchoring, confusion, or false confidence	Provide actionable information about source, process, performance, and limits
Accountability	Delegation, override, and responsibility allocation	Blame shifting, defensive rejection, excessive conformity, or misaligned discretion	Align authority, expertise, incentives, contestability, and responsibility
Fairness	Participation, acceptance, and contestation of allocation	False objectivity, biased data, biased oversight, or feedback loops	Audit fairness claims and make allocation decisions contestable
Efficiency	Use, non-use, learning, and adaptation	Underuse, overuse, dependency, skill loss, or welfare-reducing interaction	Evaluate human-AI systems dynamically, not only average technical performance

Because the reviewed studies vary in design, setting, population, and technology, the survey treats experimental evidence as mechanism-identifying rather than uniformly generalizable. Laboratory and online experiments provide clean identification of responses to disclosure, incentives, advice, delegation rights, or explanations, but often have uncertain external validity. Field experiments provide stronger evidence about

behavior in natural institutional settings, but their estimates may depend on the specific population, organization, task, platform, or selection into AI use. The findings are interpreted as evidence about the direction and conditions of calibrated or miscalibrated reliance, not as context-free average effects of AI.

The technologies covered range from predictive and allocative systems, such as classifiers, scoring tools, recommender systems, and automated hiring or lending tools, to conversational agents and generative AI systems, including large language models. The former primarily raise questions about accuracy, bias, opacity, delegation, and institutional legitimacy. The latter add interactivity, social presence, fluency, and anthropomorphic cues. Across these technologies, the central question remains the same: whether humans rely on AI in ways warranted by the system's capabilities, limitations, and safeguards.

The five domains reviewed below follow a sequence from upstream conditions to outcome domains. Privacy concerns the informational inputs of AI systems; transparency concerns the interface through which users interpret AI outputs; accountability concerns authority and responsibility for AI-mediated decisions; fairness concerns the legitimacy of AI-mediated allocation; and efficiency concerns whether human-AI systems generate welfare gains. Together, these domains identify the conditions under which reliance on AI becomes warranted.

1. Privacy, Data Disclosure, and Informational Constraints

AI systems depend on data. Before individuals can rely on AI-generated recommendations, personalized services, predictive assessments, or automated decisions, they are often asked to disclose personal information. Privacy is the first constraint on calibrated reliance because it concerns the informational inputs from which AI systems derive value. If users cannot assess what data they provide, how those data are transformed into predictions, who benefits from the resulting inferences, and who bears downstream risks, then later reliance on AI outputs is already built on an unstable informational foundation.

The economic problem is that data disclosure is both an individual choice and an input into a larger algorithmic production process. Users may receive immediate benefits from disclosure, such as personalization, access, convenience, or better recommendations, while many costs are delayed, probabilistic, opaque, or imposed on others. For this reason, disclosure behavior cannot be interpreted straightforwardly as trust. It may reflect informed confidence in a trustworthy institution, but it may also reflect cognitive limits, strategic pressure, social externalities, or interface-induced reassurance.

The following evidence does not always examine AI systems directly. Its relevance lies in identifying behavioral mechanisms of data disclosure—limited attention, valuation difficulty, strategic incentives, social externalities, and perceived control—that become central once personal data serve as inputs into AI prediction,

personalization, and allocation. Privacy evidence is not treated here as direct evidence of trust in AI, but as evidence about the upstream conditions under which later reliance on AI can become warranted or miscalibrated.

A first distortion arises from bounded rationality. The privacy paradox—strong stated privacy concerns combined with extensive disclosure—does not necessarily indicate indifference to privacy. Alashoor et al. (2022) show that privacy concerns guide disclosure less under low-effort information processing, cognitive depletion, and positive mood. Svirsky (2022) adds that individuals may avoid information about privacy consequences when learning about them would impose psychological discomfort or difficult trade-offs. In AI-mediated environments, where consent and disclosure decisions are often fast, routine, and interface-driven, such behavior should not be read as informed trust. A related problem is valuation difficulty. Lee and Weber (2025) show that many individuals exhibit stable privacy preferences in simple choices but struggle to translate them into consistent monetary valuations. Tomaino et al. (2023) further show that valuation depends on the medium of exchange: people demand higher compensation for data when paid in money than when data are exchanged for goods. Lin (2022) distinguishes intrinsic privacy preferences, in which privacy is valued as a right or expressive good, from instrumental preferences, in which privacy protects against economic loss or strategic disadvantage. These findings suggest that privacy preferences are not necessarily absent or unstable; rather, users often face difficulty articulating, valuing, and implementing them in the disclosure environments through which AI systems obtain data.

A second distortion arises from strategic incentives. Individuals do not disclose data in isolation. They disclose in markets, platforms, workplaces, and competitive environments where data sharing affects access to benefits, prices, rankings, or opportunities. Ackfeld and Güth (2023) show that willingness to share personal information increases under strategic incentives and peer comparison, suggesting that disclosure may reflect competitive pressure rather than confidence in the AI system or the institution collecting data. Heiny et al. (2025) show that data policies affect consumer strategies under behavior-based pricing: consumers may share data under open data policies to induce price competition but hide information under exclusive data policies to avoid loyalty surcharges. Fehrenbach and Herrando (2021) provide evidence consistent with a privacy calculus in real-life settings, showing that consumers distinguish positive and negative consequences of personalization. These findings imply that privacy-related trust is institutionally embedded. The same individual may appear trusting or distrustful depending on whether disclosure occurs under competitive pressure, price discrimination, personalization benefits, or credible institutional safeguards. The common mechanism is that disclosure becomes a strategic response to market design rather than a transparent expression of confidence in the data collector or the AI system.

A third distortion arises from externalities. Personal data often reveal information not only about the individual who discloses, but also about family members, peers, demographic groups, or future users. This is especially important for AI because predictive systems aggregate individual disclosures into collective inference structures. Friehe et al. (2025) show that willingness to sell personal data decreases when individuals are made aware that disclosure compromises others' privacy, especially when injunctive norms are salient. Wang et al. (2025) find that complexity and negative externalities can lead to over-sharing relative to equilibrium predictions. Klockmann et al. (2022, 2025) demonstrate that this is a critical upstream problem: when individual data contributions are perceived as small and non-pivotal, contributors ignore the negative intergenerational externalities of their choices, resulting in the production of less prosocial and potentially harmful algorithmic outcomes. Individual consent under-protects privacy when data generate collective inferences, collective benefits, or collective harms. At the same time, social motives can also increase disclosure when data use is perceived as collectively beneficial. Freddi and Wasenden (2024) show that collectivist preferences and trust in health authorities increased willingness to share sensitive location data during a pandemic. Hoyer and van Straaten (2022) show that anonymity changes participation in online rating systems: it protects privacy but may crowd out self-expression, although altruistically motivated users continue to contribute. These studies show that disclosure is not only a private preference-revelation problem; it is also a social-choice problem in which individual data decisions can create benefits and harms for others. The implication is that individual consent might be a weak governance device when algorithmic inference converts private disclosures into collective risks or benefits.

A fourth issue concerns institutional design. Because privacy choices are cognitively demanding, strategic, and socially interdependent, regulation and interface design can either improve or distort calibration. Godinho de Matos and Adjerid (2022) show in a large-scale field experiment that GDPR-compliant consent mechanisms increasing transparency and control can lead to higher data allowances. This does not necessarily mean that users become less privacy-conscious; perceived control may make disclosure feel more legitimate. The calibration question is whether such additional disclosure reflects better understanding of data consequences or merely reassurance produced by the appearance of control. Hernández et al. (2025) show that explainable AI can support privacy-preserving behavior by helping users understand how an algorithm learns about them. Transparency is most valuable when it makes inference mechanisms understandable and actionable, not when it merely induces comfort or consent.

The calibration problem can begin before an AI recommendation is ever received. Users may disclose too much because they are inattentive, strategically pressured, unaware of externalities, or reassured by control interfaces; they may disclose too little because they distrust institutions, value privacy intrinsically, or anticipate exploitation. Observed disclosure is not evidence of trust by itself. It is a (boundedly rational)

economic choice made under informational constraints, incentives, externalities, and institutional rules. Whether disclosure supports warranted reliance depends on whether users understand the data consequences of AI systems and whether institutions govern those consequences credibly.

Privacy shapes the informational inputs of AI systems. Once data have been disclosed or inferred, calibrated reliance depends on whether users and affected parties can understand how those inputs are transformed into recommendations, predictions, or decisions. This shifts attention from disclosure to transparency.

2. Transparency, Explainability, and the Informational Interface

Privacy identifies the upstream informational conditions under which AI systems are built. Transparency concerns the next step in the chain: whether users and affected parties can observe and interpret AI systems sufficiently to judge when reliance is appropriate. Even if an AI system is accurate or privacy-preserving, users rarely observe its data, objectives, error structure, or institutional safeguards directly. They encounter signals: disclosure that AI is involved, explanations of how a recommendation was generated, feedback about performance, or interface cues that make the system appear more or less human-like. Transparency is the informational interface through which users form beliefs about whether reliance is warranted.

This makes transparency an economic problem of information design. More information is not necessarily better. Information improves calibration only when it helps users distinguish reliable from unreliable systems, understand relevant limitations, compare algorithmic and human judgment, or identify when contestation is appropriate. It can undermine calibration when it triggers source-based aversion, creates confusion, anchors users on inaccurate scores, or increases confidence without increasing actual reliability. The relevant question is not whether transparency increases trust, but whether it improves the match between trust and trustworthiness. The evidence reviewed here distinguishes four forms of transparency: source disclosure, which reveals whether AI is involved; process information, which explains how an output was generated; performance information, which helps users compare AI and human accuracy; and interface information, which shapes perceptions through voice, timing, social presence, or anthropomorphic cues.

A first form of transparency is source disclosure: users or affected parties are told whether AI is involved in producing, mediating, or evaluating an outcome. Keppeler (2024) shows that job candidates express less interest in public-sector job offers when they are told AI was used to identify them. Friedrichsen et al. (2026) find that listeners value AI-generated music as much as human-made music when the origin is unknown, but appreciation and willingness to pay decline sharply after AI disclosure. In a randomized field experiment with a large truck-sharing platform, Xu et al. (2026) demonstrate that explicitly disclosing a voice chatbot's AI identity at the beginning of a conversation results in an 11% reduction in driver response rates and a significant

decline in order acceptance intention. Irlenbusch et al. (2026) find that AI involvement in hiring evaluations reduces application rates, especially among low-competitive women. These studies show that AI disclosure can affect behavior, but not through a single mechanism. In some settings, disclosure changes evaluation of a fixed output; in others, it changes expectations about fairness, accountability, privacy, or interaction quality. Across these cases, the AI label works less like a neutral fact and more like an institutional signal whose meaning depends on the task, the affected party, and the perceived stakes.

The calibration implication is ambiguous. Disclosure may appropriately reduce reliance when AI involvement signals legitimate concerns about fairness, authenticity, surveillance, or accountability. But it may also produce unwarranted aversion to useful systems if users treat "AI" as a negative source cue independent of performance. Conversely, source disclosure may be insufficient to prevent over-reliance. Leib et al. (2024) show that people follow dishonesty-promoting advice from AI to a similar extent regardless of whether the source is disclosed. Source transparency may be important for legitimacy, but it is insufficient for calibration: it tells users who or what produced an output, but not whether the output deserves reliance.

A second form of transparency is process information: users receive explanations of how an AI system reaches a decision, what features it uses, or why it produced a recommendation. Process transparency is often treated as a remedy for algorithm aversion, but the experimental evidence suggests that explanations help only when they are interpretable and decision-relevant. In a movie-review classification task, Schmidt et al. (2020) show that transparency features such as keyword highlighting or confidence scores can reduce human trust in the system when the model's internal logic appears unintuitive from a human perspective. Dargnies et al. (2026) similarly find that explaining how a hiring algorithm weights applicant characteristics does not increase acceptance. These results suggest that resistance to AI may reflect overconfidence, control preferences, or institutional skepticism rather than a simple lack of information.

A third form of transparency is performance information. Users need to know not only that AI is involved or how it works, but when it performs better or worse than relevant alternatives. Feedback and incentives can improve calibration more effectively than one-shot disclosure or technical explanation. Filiz et al. (2021) show that experience and feedback reduce algorithm aversion in forecasting tasks as subjects learn about relative performance. The governance implication is that users should receive task-relevant evidence about AI performance relative to realistic alternatives, including human judgment.

A fourth form of transparency is interface information. Users do not encounter AI systems only through formal disclosures and explanations; they also respond to voice, timing, social presence, anthropomorphic cues, and the stage of interaction. Xu et al. (2026) show that anthropomorphic cues in voice chatbot design, such as vocal fillers, can mitigate the negative effect of identity disclosure by increasing perceived social presence. Adam et

al. (2022) show that customers prefer humans at early sales stages, where social presence matters, but prefer automated agents when providing contact information, where efficiency and lower effort expectancy dominate. Using incentivized auctions for artworks, Lane et al. (2025) demonstrate that consumers are generally indifferent toward the creator's identity, exhibiting nearly identical willingness to pay for both human- and AI-generated art. This aggregate indifference, however, masks a deep polarization: individuals with positive views of technology pay a premium for AI-produced work when its origin is disclosed, while those hostile to AI show a significantly stronger preference for human labor. These findings show that reliance is task-, stage-, and user-specific. They also highlight a calibration risk: interface cues can make AI feel more socially competent or easier to use without providing evidence that its outputs are more reliable, fair, or accountable. The mechanism is not transparency in a narrow disclosure sense, but cue-based belief formation at the human-AI interface.

Transparency supports calibration only when it helps users distinguish between reliable, unreliable, and contestable AI outputs. It can support warranted reliance when it helps users identify AI involvement, understand relevant limitations, compare performance, or know when contestation is appropriate. But transparency can also miscalibrate reliance: labels may trigger source-based aversion, explanations may confuse, disclosed scores may anchor behavior, anthropomorphic cues may create false confidence, and performance feedback may be absent or hard to interpret. Transparency should be evaluated not by whether it increases trust, but by whether it improves users' ability to distinguish when AI deserves reliance, refusal, or challenge.

Transparency can help users judge AI outputs, but information alone does not determine how those outputs are used. Reliance also depends on who has authority to accept, reject, override, or delegate to AI, and who bears responsibility for the consequences. This makes accountability the next condition for calibrated reliance.

3. Accountability, Delegation, and Responsibility

Transparency determines what users and affected parties can observe about AI systems. Accountability determines who is authorized to act on that information and who is answerable for the consequences. This is the third constraint on calibrated reliance. Even when users know that AI is involved, understand something about its logic, or receive evidence about its performance, reliance may still be miscalibrated if decision rights, incentives, and responsibility are misaligned. Humans may under-rely on AI because they fear being blamed for algorithmic errors or do not want to relinquish control. They may over-rely on AI because delegation creates moral distance, obscures intent, or provides a shield against punishment. Warranted reliance requires not only information about AI systems but also institutional rules that align authority with responsibility. The evidence

reviewed below points to three recurring accountability problems: delegation can be used to shift responsibility, decision authority can be misallocated between humans and AI, and accountability pressures can lead experts either to reject useful AI or to conform too strongly to it.

A first accountability problem is strategic delegation. Humans may rely on AI not because they expect better decisions, but because delegation creates moral distance, obscures intent, or reduces exposure to blame. AI can change the moral and strategic structure of delegation by separating intention, instruction, training, and outcome. Hüholt and Szech (2026) show that responsibility shifting can make AI attractive even in morally sensitive decisions. In two experiments with real donation consequences, participants delegated moral choices more often to AI than to another human, suggesting that algorithm aversion in moral domains can reverse when AI provides an opportunity to offload responsibility. Köbis et al. (2025) show that delegation to AI can increase dishonest behavior: principals are more likely to request dishonesty through AI with supervised learning or goal-setting interfaces because these indirect forms of delegation create ambiguity about intent. Feier et al. (2022) provide related evidence on punishment avoidance: in their experiment, delegators were treated more leniently after bad monetary outcomes when failure resulted from a machine agent, while delegation to a human agent did not produce a comparable reduction in punishment. In these settings, reliance on AI is not necessarily motivated by expected accuracy or efficiency. It may be motivated by the opportunity to benefit from harmful outcomes while reducing personal exposure to blame.

Accountability becomes more complex as AI systems become more agentic and interactive. Johnson and Obradovich (2024) show that AI agents can display behavior consistent with trust-like responses to humans when machine incentives are at stake. For the present argument, the implication is not that machines “trust” in the human sense, but that accountability frameworks may increasingly need to govern reciprocal human-AI interaction rather than one-sided human use of algorithmic tools.

A second accountability problem concerns decision authority. Even when AI advice is available and transparent, outcomes depend on who has the right to decide, override, or delegate. Ivanova-Stenzel (2025) argues that algorithm aversion may stem from reluctance to relinquish decision authority rather than distrust of algorithms specifically: participants underuse both human and AI agents even when those agents are demonstrably superior. Ivanova-Stenzel and Tolksdorf (2024) find that performance feedback increases reliance on AI but rarely eliminates the desire for personal control. These studies show that control has value independent of expected performance. If individuals remain answerable for outcomes, they may reasonably hesitate to cede authority to systems they do not fully control.

The allocation of authority is domain-specific. In organizations, Kim et al. (2024) show that managers use their authority to override algorithmic recommendations in pursuit of secondary objectives, like prior beliefs or

conflicts of interest, thereby reducing predictive gains. Kawaguchi (2021) shows that agency costs and conflicts of interest can prevent retail workers from following algorithmic advice. In these cases, human discretion can dilute the value of AI when incentives are misaligned. Greiner et al. (2026) provide complementary evidence that organizational control systems can also increase reliance on algorithmic advice. They show that individual performance incentives and tournament incentives increase reliance on algorithmic advice relative to fixed pay. Framing an algorithm as "human-augmented" increases use under fixed pay, suggesting that people are more receptive when AI is presented as complementing rather than replacing human judgment. Human authority can also generate unique social value; Bauer and Gill (2024) highlight the strategic complexity of hybrid systems by showing that when a human decision-maker overrides a disclosed algorithmic score, the affected individual perceives this as a genuine act of kindness. Bianchi and Brière (2026) find that robo-advising works best when investors retain freedom to follow or ignore advice, because autonomy keeps them engaged. Krakowski et al. (2026) show that authority structures should be matched to cognitive styles in unstructured tasks. These findings reject a simple rule that more automation or more human control is always better. Calibrated reliance requires matching decision rights to task structure, incentives, expertise, and responsibility. The shared lesson is that decision authority is itself a design variable: the efficient allocation of control depends on whether human discretion adds information, motivation, legitimacy, or merely noise and agency costs.

A third accountability problem concerns expertise under asymmetric responsibility. Experts may resist AI not because they are irrationally averse to algorithms, but because they expect to be held responsible for mistakes produced by systems they do not fully control. Allen and Choudhury (2022) show that experienced workers exhibit greater algorithm aversion because they feel more accountable for unintended consequences of accepting inaccurate advice. What appears as under-reliance may be a rational response to asymmetric accountability: experts bear responsibility without fully controlling or understanding the AI system.

At the same time, disagreement with AI can create value when humans possess relevant contextual information. Wang et al. (2026) show that humans improve loan evaluations when they successfully correct bad algorithmic recommendations, especially in high-information environments where they can interrogate the advice. The goal is neither maximum compliance nor unrestricted discretion, but justified reliance and justified disagreement.

A final accountability problem concerns aggregation. Individual reliance decisions can affect collective information and system-level performance. Fügener et al. (2021) show that human-AI collaboration can improve individual accuracy while reducing diversity in human judgment, thereby weakening crowd wisdom unless advice is personalized to preserve independent information. Chevrier et al. (2024a) find that delegation is driven primarily by expected payoffs rather than error frequency or magnitude, documenting algorithmic

appreciation based on relative performance comparisons. These findings show that accountability must consider not only whether a single user should follow AI advice, but also whether institutional incentives encourage excessive convergence, defensive rejection, or productive diversity in human judgment.

The calibration problem is also a problem of governance. Privacy determines how data enter AI systems; transparency determines what users and affected parties can observe; accountability determines how authority, discretion, and responsibility are allocated. Miscalibration arises when people delegate to AI to evade blame, reject useful AI because they bear responsibility without control, override algorithms for misaligned objectives, or converge too strongly on common AI advice. Warranted reliance requires institutional arrangements that align decision rights with expertise, incentives, contestability, and responsibility.

Accountability structures determine how AI-mediated decisions are made and who is answerable for them. The next question is whether those decisions allocate opportunities in ways that affected parties can regard as legitimate, unbiased, and contestable. This moves the analysis from governance conditions to fairness outcomes.

4. Fairness, Bias, and Institutional Constraints

Privacy, transparency, and accountability identify governance conditions under which reliance on AI can be warranted. Fairness concerns the first outcome domain: how AI systems transform data, decision rules, and organizational authority into predictions, rankings, recommendations, and decisions that allocate opportunities. Once AI systems screen applicants, recommend jobs, evaluate workers, assist teachers, guide lending decisions, or allocate access to services, trust depends on whether affected parties regard algorithmic allocation as legitimate, unbiased, and contestable. Fairness is the fourth constraint on calibrated reliance. The evidence reviewed below distinguishes four fairness margins: participation, decision rules, organizational control, and longer-run distributional effects.

This makes trust in AI partly a form of institutional trust. Applicants, workers, managers, students, and borrowers rarely observe an algorithm's training data, objective function, or validation procedure directly. They infer from the deploying organization whether the system is likely to reduce human bias, reproduce historical inequality, or legitimize contested decisions under a veneer of objectivity. AI may be trusted because it appears less biased than humans, rejected because it is opaque or rigid, or contested because it automates institutional disadvantage. The relevant economic question is not whether algorithms are trusted more or less than humans in general, but whether algorithmic allocation improves the match between decision quality, fairness, and procedural legitimacy.

A first fairness margin is participation: whether AI changes who enters an allocation process. If affected parties expect human evaluators to be biased, algorithmic assessment may increase willingness to enter an allocation process. Avery et al. (2024) show in a technology recruitment field experiment that informing candidates they would be evaluated by AI increased application completion among women by about 30 percentage points relative to men, closing a gender gap in completion. Pethig and Kroenung (2023) similarly show that women are more likely to choose algorithmic evaluation when the alternative is a male human evaluator, with perceived algorithmic objectivity mediating this preference. These findings show that trust in AI may be comparative: the algorithm is not necessarily trusted because it is viewed as intrinsically fair, but because the human alternative is distrusted. The calibration issue is whether this comparative trust is warranted. Higher participation improves welfare only if algorithmic assessment actually reduces bias or improves procedural legitimacy; otherwise, AI may attract applicants by signaling objectivity without delivering fairer outcomes. The mechanism is selection into the allocation process: AI can change who applies, participates, or opts out before any formal decision is made.

This participation effect is economically important but fragile. If applicants infer fairness merely from automation, reliance may be misplaced. Awad et al. (2025) and Ip (2025) show that standard AI assessment does not necessarily increase gender diversity; the effects on diversity depend on specific debiasing methods, such as equality of prediction, that attract qualified female applicants without lowering applicant quality. Fumagalli et al. (2022) find that high-performing workers are more likely to prefer algorithmic recruitment because they perceive it as more meritocratic and less error-prone. Together, these studies show that fairness affects selection into AI-mediated environments. Algorithmic systems can broaden participation when they credibly reduce expected bias, but they can also create false reassurance if objectivity is inferred from automation alone.

A second fairness margin concerns decision rules: whether AI reduces, reproduces, or transforms bias. AI can mitigate human bias by standardizing evaluation and reducing the influence of stereotypes, but it can also reproduce bias when trained on biased data or optimized for inappropriate objectives. Pisanelli (2022) reports that replacing human recruiters with automated resume screening substantially reduced a gender gap in interview invitations. Avery et al. (2024) likewise find that AI-generated scores eliminated gender gaps in assessments even when applicant gender was known. These findings suggest that algorithmic tools can discipline noisy or biased human judgment. Yet Hu et al. (2026) show in microlending that human preference-based and belief-based biases can carry over into machine learning algorithms. Cowgill et al. (2020) similarly emphasize that biased predictions often arise from biased training data rather than biased programmers. Zhang and Kuhn (2024) warn that job recommender systems can steer applicants toward gender-stereotypical roles when declared gender is used as a direct matching input. Fairness cannot be inferred from automation

itself. It depends on how training data are generated, which objective function is optimized, which fairness criterion is imposed, and whether the deploying institution has incentives to detect and correct biased outcomes.

A third fairness margin concerns organizational control: whether human oversight corrects or reintroduces bias. Even a well-designed algorithm affects outcomes only through organizational use: managers may delegate to it, ignore it, override it, or combine it with human judgment. Dargnies et al. (2026) show that managers often fail to delegate hiring decisions to more efficient algorithms because they overestimate their recruitment abilities; feedback on their own errors increases delegation. Awuah et al. (2026) find in teacher screening in Ghana that full automation outperformed human-AI assistance because human evaluators ignored AI recommendations and reverted to idiosyncratic grading. These findings caution against treating human oversight as a fairness safeguard by default. Oversight can correct algorithmic errors only when humans have further relevant information, incentives to use it, and authority to challenge the system; otherwise, discretion may reintroduce bias, noise, or overconfidence. The mechanism is implementation fidelity: even accurate and less biased tools can fail when human users have incentives or beliefs that lead them to ignore, distort, or selectively apply algorithmic recommendations.

The control margin also depends on organizational context. Dargnies et al. (2026) find that explaining how a hiring algorithm weights applicant characteristics does not necessarily increase acceptance, suggesting that resistance may reflect overconfidence, control preferences, or skepticism about institutional motives rather than a simple lack of information. Li et al. (2026) find that lower-class employees perceive greater risks in using large language models (LLMs) for workplace help-seeking, indicating that adoption and trust are structured by workplace hierarchy. Glickman and Sharot (2025) warn that feedback loops between humans and biased AI can amplify human bias, and Xu et al. (2025) show that LLMs may exhibit self-preference bias by favoring resumes that resemble their own generative style. This suggests a new fairness risk specific to generative AI: models may reward linguistic or stylistic proximity to AI-generated text, thereby changing which applicants appear qualified. These studies connect fairness to organizational hierarchy and feedback dynamics, showing that algorithmic bias is not only a property of models but also of the social systems in which models are embedded.

A fourth fairness margin concerns longer-run distributional effects: who gains skills, access, and labor-market rewards from AI. AI may affect not only individual allocation decisions, but also who acquires skills, whose performance improves, and which signals are rewarded in labor markets. Beyond decision-level effects, Bao et al. (2026) show that AI trainers improved Go outcomes for all students and reduced the gender gap, plausibly by providing an emotionally neutral and interactive learning environment. AI may expand opportunity when it

reduces biased instruction, supplies scalable support, or rewards scarce complementary skills. It may deepen inequality if access to AI tools or AI-related credentials becomes a new gatekeeping mechanism.

Fairness shifts the calibration question from whether AI is used to whether AI-mediated allocation is legitimate, unbiased, and contestable. Miscalibration can occur when applicants over-trust automation because it appears objective, when managers underuse beneficial algorithms because of overconfidence, when organizations rely on biased data or underspecified objectives, or when AI systems create feedback loops and new distributional inequalities. Trust in AI is warranted only when fairness claims are empirically credible, organizationally enforced, and open to challenge.

Fairness concerns the legitimacy of AI-mediated allocation. But legitimate allocation is only one dimension of economic value. The final question is whether human-AI systems improve productivity, learning, decision quality, and welfare over time. This is the efficiency test of calibrated reliance.

5. Efficiency, Effectiveness, and Welfare

The preceding sections identify the conditions and outcomes that determine whether reliance on AI is warranted. Efficiency is the final outcome domain because it asks whether governed human-AI systems generate welfare gains. This is not simply a question of technical performance or average productivity. AI may improve prediction, production, learning, or allocation in one setting while reducing welfare in another through over-reliance, skill loss, weakened cooperation, or misaligned incentives. Whether efficiency gains materialize depends on whether reliance is calibrated in practice.

Efficiency should not be treated as separate from trust. Under-reliance wastes useful AI: individuals and organizations fail to use systems that could improve decisions, reduce judgment noise, or expand expertise. Over-reliance creates the opposite problem: users follow deficient advice, become dependent, lose skills, reduce cooperation, or respond to persuasive but unreliable systems. The relevant welfare question is not whether AI improves performance on average, but when human-AI systems outperform the available alternatives, for whom, and over what time horizon.

A first welfare margin is production. Evidence from workplaces and markets shows that AI can generate substantial productivity gains when its comparative advantage is clear and the task environment supports effective use. Brynjolfsson et al. (2025) find that a generative AI conversational assistant increased customer support productivity by 15 percent, with gains concentrated among less experienced and lower-skilled workers. Zhang and Narayandas (2026) show that AI assistance in online chats reduced response times and improved customer sentiment. Kanazawa et al. (2026) find that AI demand-forecasting tools reduced taxi drivers' cruising time by 5.3 percent, especially for less-experienced drivers. Jabarian and Henkel (2026) show

that AI voice agents in recruiting improved match quality and retention by reducing information-collection variance relative to human recruiters. These studies show how AI can create welfare gains by standardizing information processing, reducing experience gaps, and making scarce expertise scalable.

These returns are not automatic. Firpo, Niemann, and Danilov (2025) find limited average labor-market returns to AI qualifications on resumes, suggesting that AI-related skills generate economic value only when they are scarce, credible, and relevant to the task environment. Strobel (2025) shows that automated bonus evaluations can harm worker performance when workers misinterpret how procedural rigidity affects performance thresholds. These findings reinforce the central calibration point: AI adoption creates value only when users understand the task, incentives, and institutional rules under which the technology operates.

A second welfare margin is matching and interaction. Human-AI collaboration creates value only when tasks are allocated to the agent—human, algorithmic, or hybrid—best suited to perform them. Fügener et al. (2022) show that human-AI collaboration can outperform AI alone, but only when AI delegates selectively to humans; when humans must decide when to use AI, they struggle because they cannot accurately identify which tasks are difficult for them. Bayer and Renou (2026) provide complementary evidence that the welfare effects of human-AI collaboration depend on task difficulty and the type of reasoning required. In an interactive reasoning task, they find that participants perform better with other humans on simple tasks, but better with AI or expert human counterparts on difficult tasks. The expert condition shows that this difference is driven not by whether the counterpart is human or nonhuman, but by whether participants know that the counterpart reasons correctly. This suggests that AI can improve performance in complex reasoning environments by reducing certain forms of strategic uncertainty, but may underperform human interaction when social reasoning and perspective-taking are sufficient for the task. Luo et al. (2021) find an inverted-U pattern in AI coaching for sales agents: middle-ranked agents benefit most, bottom-ranked agents suffer from information overload, and top-ranked agents show stronger aversion. Dvorak et al. (2025) show that AI-mediated interaction can reduce welfare in economic games when humans know they are interacting with ChatGPT, leading to lower cooperation and fairness. Opitz et al. (2025) show that machine learning can outperform the best human-designed incentive scheme by assigning incentives to workers based on predicted individual responses. Relatedly, Dargnies et al. (2025) show that AI-supported hiring prediction improves when behavioral and cognitive measures are elicited directly from applicants. This suggests that AI can create efficiency gains not only by processing existing data, but also by changing what information is collected. At the same time, the result links efficiency back to privacy and fairness, because richer applicant measurement increases informational demands and raises questions about contestability in AI-mediated selection. This illustrates both the efficiency promise and the governance risk of AI: individualized choice architecture can improve performance while making the basis of treatment harder for workers to understand or contest. These

findings show that welfare depends not simply on making AI available, but on matching AI support to task difficulty, user capability, social context, and institutional safeguards.

A third welfare margin is belief calibration. AI creates value only if users can calibrate beliefs about advice quality, uncertainty, and the limits of model competence. Jung and Seiter (2021) show that calibrated reliance on algorithmic advice depends on the decision environment in which users evaluate their own judgment. Their main result is that time pressure reduces algorithm aversion. Under time pressure, participants become less confident in their own forecasts and are more willing to rely on the algorithm. Diecidue et al. (2026) find that decision-makers can fail to differentiate adequately between changes in the probability that machine predictions are accurate. Germann and Merkle (2023) find no general algorithm aversion in delegated investing: investors care more about returns than source, but struggle to distinguish skill from luck, slowing migration toward better-performing algorithms. Klingbeil, Grützner, and Schreck (2024) show that users may over-rely on AI advice, even when it contradicts the available context. Chevrier et al. (2024b) introduce the concept of algorithmic credulity, showing that participants are more likely to follow deficient advice when it comes from an algorithm than from a human, thereby lowering earnings. Cheng et al. (2026) identify another form of over-reliance: sycophantic AI increases users' conviction that they are right and reduces intentions to repair interpersonal conflicts. These studies show that welfare losses can arise not only from algorithmic error, but from users' inability to interpret reliability, uncertainty, and socially persuasive cues.

A fourth welfare margin is dynamic human capital. AI may improve immediate performance while changing the skills, beliefs, and adaptive capacities on which future performance depends. Fischer et al. (2025) show that unrestricted access to an AI tutor increased student test performance by 0.23 standard deviations without crowding out independent reading effort, suggesting that AI can complement rather than replace learning effort. Hausman et al. (2025) find that AI availability increased grades in AI-compatible higher-education courses, but also raise concerns that over-reliance may create knowledge gaps. These findings show that the welfare effects of AI depend not only on immediate output gains, but also on whether users continue to acquire the expertise needed to evaluate, contest, and improve AI-supported decisions. Dynamic effects also arise through occupational expectations and adaptation. Rostam-Afschar et al. (2025) find that German tax advisors initially underestimated task automatability; providing accurate information increased adoption intentions and expectations about future work, although it did not immediately change hiring plans. This evidence suggests that AI can affect welfare before observable productivity or employment changes occur, by reshaping beliefs about future tasks, skills, and organizational adjustment. The welfare test of AI is intertemporal. Short-run performance gains are beneficial when they complement effort, learning, and adaptation; they are more ambiguous when they substitute for understanding, create dependency, or delay necessary adjustment.

The welfare effects of AI depend on whether reliance is calibrated over time. Privacy, transparency, accountability, and fairness are not separate normative constraints to be balanced against performance after the fact. They are part of the production function of AI's economic value. When these constraints are well designed, AI can increase productivity, improve learning, reduce judgment noise, and expand access to expertise. When poorly designed, the same technology can produce underuse, overuse, biased allocation, misplaced authority, dependency, weakened human expertise, and welfare-reducing interaction.

Conclusion

The experimental evidence reviewed in this survey suggests that human trust in AI is best understood not as a general attitude toward machines, but as a pattern of reliance shaped by information, incentives, institutions, and responsibilities. People sometimes reject useful AI because they distrust opaque systems, value autonomy, anticipate unfairness, or question the motives of the deploying institution. In other settings, they over-rely on deficient AI, follow misleading or unethical advice, disclose excessive data, or use algorithmic delegation to diffuse responsibility. The central problem is not whether humans trust AI too much or too little, but whether reliance is calibrated to what the system can do, what risks it creates, and what safeguards surround its use.

The five domains reviewed here identify complementary sources of miscalibration. Privacy evidence shows why disclosure cannot be treated as informed trust. Transparency evidence shows that disclosure, explanation, performance information, and interface cues can either improve calibration or generate aversion, anchoring, and false confidence. Accountability evidence shows that reliance depends on how authority, discretion, and responsibility are allocated. Fairness evidence shows that AI-mediated allocation is trustworthy only when data, objectives, debiasing procedures, and organizational incentives support legitimate and contestable outcomes. Efficiency evidence shows why these constraints matter economically: AI creates value when it improves productivity, learning, and decision quality, but can reduce value through dependency, credulity, skill loss, or welfare-reducing interaction. Taken together, these findings suggest that AI systems should be evaluated as sociotechnical choice architectures rather than as isolated technical tools.

This perspective has direct implications for governance and design. The aim should not be to increase trust as such. More trust is harmful if it leads to over-reliance, excessive disclosure, credulity, or responsibility diffusion; less trust is harmful if it prevents the use of systems that improve decisions, reduce bias, or support learning. The appropriate objective is warranted trust: reliance calibrated to the system's actual capabilities, limitations, and safeguards. The evidence reviewed here gives this governance objective more concrete content. It suggests five complementary governance levers: making data consequences intelligible, providing actionable

transparency, aligning authority with responsibility, auditing fairness claims, and evaluating welfare effects over time.

Experimental economics contributes to AI governance by identifying behavioral mechanisms that governance must anticipate. It shows that disclosure, explanation, delegation, oversight, and AI advice do not operate mechanically: their effects depend on cognitive limits, incentives, perceived control, institutional motives, responsibility allocation, and social spillovers. These findings do not yet provide complete institutional equilibria. Their value for governance is diagnostic: they identify the incentives, information frictions, and responsibility gaps that regulation, organizational design, and interface design must address.

The rise of generative AI makes this governance agenda more urgent. Large language models increasingly mediate communication, education, hiring, advice, workplace support, and social interaction, while their fluency, adaptability, and social responsiveness can obscure errors, biases, sycophancy, and institutional interests. Experimental economics is well suited to study these risks because it can vary rules, incentives, information, delegation rights, and accountability structures while observing how people update beliefs, rely on advice, shift responsibility, and respond strategically. Future research should move beyond binary comparisons between humans and algorithms toward experiments on governable human-AI systems: hybrid teams, sequential delegation, adaptive explanations, strategic AI agents, justified disagreement, audit and contestation procedures, liability rules, and regulatory environments that shape calibrated reliance over time.

Trust in AI is neither algorithm aversion nor algorithm appreciation. It is a form of reliance under vulnerability, and its value depends on whether that reliance is warranted by the system's capabilities, limits, and institutional safeguards. The task for research, design, and governance is not to maximize trust, but to structure AI-mediated environments so that humans use beneficial systems, question unreliable ones, and remain able to contest decisions whose outputs, objectives, or institutional contexts do not deserve trust.

References

Ackfeld, V., & Güth, W. (2023). Personal information disclosure under competition for benefits: Is sharing caring? Games and Economic Behavior, 140, 1–32.

Adam, M., Roethke, K., & Benlian, A. (2022). Human versus automated sales agents: How and why customer responses shift across sales stages. Information Systems Research.

Alashoor, T., Keil, M., Smith, H. J., & McConnell, A. R. (2022). Too tired and in too good of a mood to worry about privacy: Explaining the privacy paradox through the lens of effort level in information processing. Information Systems Research.

Alekseev, A., & Strobel, C. (2026). A taxonomy of AI experiments. Journal of Behavioral and Experimental Economics, 121, 102525.

Allen, R. T., & Choudhury, P. (2022). Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. Organization Science, 33(1), 149–169.

Aquilino, L., Di Dio, C., Manzi, F., Massaro, D., Bisconti, P., & Marchetti, A. (2025). Decoding trust in artificial intelligence: A systematic review of quantitative measures and related variables. Informatics, 12(3), 70.

Avery, M., Leibbrandt, A., & Vecci, J. (2024). Does artificial intelligence help or hurt gender diversity? Evidence from two field experiments on recruitment in tech (CESifo Working Paper No. 10996).

Awad, E., Balafoutas, L., Chen, L., Ip, E., & Vecci, J. (2025). Artificial intelligence and debiasing in hiring: Impact on applicant quality and gender diversity. SSRN.

Awuah, K., Krenk, U., & Yanagizawa-Drott, D. (2026). Augment or automate? An early field experiment with generative AI in hiring. University of Zurich.

Bacha, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2024). A systematic literature review of user trust in AI-enabled systems: An HCI perspective. International Journal of Human–Computer Interaction, 40(5), 1251–1266.

Bao, L., Huang, D., & Lin, C. (2026). Can artificial intelligence improve gender equality? Evidence from a natural experiment. Management Science, 72(1), 474–494.

Bauer, K., & Gill, A. (2024). Mirror, mirror on the wall: Algorithmic assessments, transparency, and self-fulfilling prophecies. Information Systems Research, 35(1), 226–248.

Bayer, R.-C., & Renou, L. (2026). Interacting with man or machine: When do humans reason better? Management Science, 72(1), 594–608. https://doi.org/10.1287/mnsc.2023.03315

Benk, M., Kerstan, S., von Wangenheim, F., & Ferrario, A. (2025). Twenty-four years of empirical research on trust in AI: A bibliometric review of trends, overlooked issues, and future directions. AI & SOCIETY, 40, 2083–2106.

Bianchi, M., & Brière, M. (2026). Human-robot interactions in investment decisions. Management Science, 72(1), 14–31.

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. The Quarterly Journal of Economics, 140(2), 889–942.

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making, 33(2), 220–239.

Caro, F., Colliard, J.-E., Katok, E., Ockenfels, A., Stier-Moses, N., Tucker, C., & Wu, D. J. (2026). Introduction to the special issue on the human-algorithm connection. Management Science, 72(1), 1–13.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J.-F., Brañas-Garza, P., Butera, L., Douglas, K. M., Everett, J. A. C., Gigerenzer, G., Greenhow, C., Hashimoto, D. A., Holt-Lunstad, J., Jetten, J., Johnson, S., Kunz, W. H., Longoni, C., Lunn, P., Natale, S., Paluch, S., Rahwan, I., Selwyn, N., Singh, V., Suri, S., Sutcliffe, J., Tomlinson, J., van der Linden, S., Van Lange, P. A. M., Wall, F., Van Bavel, J. J., & Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. PNAS Nexus, 3, 191. https://doi.org/10.1093/pnasnexus/pgae191

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391, eaec8352.

Chevrier, M., Fidanoski, F., & Teixeira, V. (2024a). Choosing between algorithmic forecasters: What drives delegation? Working Paper.

Chevrier, M., Corgnet, B., Guerci, E., & Rosaz, J. (2024b). Algorithm credulity: Human and algorithmic advice in prediction experiments (GREDEG Working Paper No. 2024-03). Université Côte d'Azur.

Chugunova, M., & Sele, D. (2022). We and It: An interdisciplinary review of the experimental evidence on how humans interact with machines. Journal of Behavioral and Experimental Economics, 99, 101897.

Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N., & Chaintreau, A. (2020). Biased programmers? Or biased data? A field experiment in operationalizing AI ethics. In: Proceedings of the 21st ACM Conference on Economics and Computation, 679-681.

Dang, Q., & Li, G. (2026). Unveiling trust in AI: The interplay of antecedents, consequences, and cultural dynamics. AI & SOCIETY, 41, 669–692.

Dargnies, M.-P., Hakimov, R., & Kübler, D. (2025). Behavioral measures improve AI hiring: A field experiment (Discussion Paper No. 532). Collaborative Research Center Transregio 190 – Rationality and Competition, Ludwig-Maximilians-Universität München & Humboldt-Universität zu Berlin.

Dargnies, M.-P., Hakimov, R., & Kübler, D. (2026). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. Management Science, 72(1), 285–301.

Diecidue, E., Gueciloueur, A., & Xia, Q. (2026). Trusting human versus machine predictions as a decision under ambiguity. Journal of Risk and Uncertainty. Advance online publication.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126.

Duran, J. M., & Pozzi, G. (2025). Trust and trustworthiness in AI. Philosophy & Technology, 38(1), 16.

Dvorak, F., Stumpf, R., Fehrler, S., & Fischbacher, U. (2025). Adverse reactions to the use of large language models in social interactions. PNAS Nexus, 4(4), pgaf112.

Fehrenbach, D., & Herrando, C. (2021). The effect of customer-perceived value when paying for a product with personal data: A real-life experimental study. Journal of Business Research, 137, 222–232.

Feier, T., Gogoll, J., & Uhl, M. (2022). Hiding behind machines: Artificial agents may help to evade punishment. Science and Engineering Ethics, 28(19).

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021). Reducing algorithm aversion through experience. Journal of Behavioral and Experimental Finance, 31, 100524.

Firpo, T., Niemann, L. & Danilov, A. (2025). The elusive returns to AI skills: Evidence from a field experiment. Rationality & Competition CRC TRR 190, Discussion Paper 552, School of Business and Economics, Humboldt-Universität zu Berlin.

Fischer, M., Rau, H. A., & Rilke, R. M. (2025). AI tutoring enhances student learning without crowding out reading effort (BiB. Working Paper 10/2025). Federal Institute for Population Research.

Freddi, E., & Wasenden, O. C. (2024). Privacy during pandemics: Attitudes to public use of personal data. Journal of Behavioral and Experimental Economics, 113, 102304.

Friedrichsen, J., Schwarz, J., & Clement, M. (2026). When music is made by AI: Effects on preferences and willingness to pay. CESifo Working Paper, 12405.

Friehe, T., Gerhards, L., & Weber, F. (2025). Keep them out of it! How information externalities affect the willingness to sell personal data online. Journal of Economic Psychology, 109, 102830.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. MIS Quarterly, 45(3), 1527–1556.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. Information Systems Research, 33(2), 678–696.

Fumagalli, E., Rezaei, S., & Salomons, A. (2022). OK computer: Worker perceptions of algorithmic recruitment. Research Policy, 51(2), 104420.

Germann, M., & Merkle, C. (2023). Algorithm aversion in delegated investing. Journal of Business Economics, 93, 1691–1727.

Glickman, M., Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. Nature Human Behaviour, 9, 345–359. https://doi.org/10.1038/s41562-024-02077-2

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660.

Godinho de Matos, M., & Adjerid, I. (2022). Consumer consent and firm targeting after GDPR: The case of a large telecom provider. Management Science, 68(5), 3330–3378.

Greiner, B., Grünwald, P., Lindner, T., Lintner, G., & Wiernsperger, M. (2026). Incentives, framing, and reliance on algorithmic advice: An experimental study. Management Science, 72(1), 302–322.

Hahne, P.-Z., & Schmoelz, A. (2026). Trusting the machine: A digital humanist perspective on misplaced trust in artificial intelligence. AI and Ethics, 6, Article 115. https://doi.org/10.1007/s43681-025-00923-1

Hannon, O., Ciriello, R., & Gal, U. (2024). Just because we can, doesn't mean we should: Algorithm aversion as a principled resistance. Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS), 6076–6085.

Hausman, N., Rigbi, O., & Weisburd, S. (2025). Generative AI's impact on student achievement and implications for worker productivity. CESifo Working Paper No. 11843.

Heiny, F., Li, T., & Tolksdorf, M. (2025). We value your privacy: Behavior-based pricing under endogenous privacy. Journal of Economics & Management Strategy.

Henrique, B. M., & Santos Jr., E. (2024). Trust in artificial intelligence: Literature review and main path analysis. Computers in Human Behavior: Artificial Humans, 2, 100043.

Hernández, P., Morales, A. J., Neeman, Z., & Pavía, J. M. (2025). Exploring the privacy paradox: An experimental investigation of privacy-preserving behavioural responses in online shopping. Journal of Behavioral and Experimental Economics, 121, 102504.

Hoyer, B., & van Straaten, D. (2022). Anonymity and self-expression in online rating systems—An experimental analysis. Journal of Behavioral and Experimental Economics, 98, 101869.

Hu, X., Huang, Y., Li, B., & Lu, T. (2026). Human-algorithmic bias: Source, evolution, and impact. Management Science, 72(1), 495–514.

Hüholt, N., & Szech, N. (2026). Trusting machines with morality—Delegating moral decisions to AI. European Economic Review, 184, Article 105255. https://doi.org/10.1016/j.euroecorev.2025.105255

Ip, E. (2025). Fair AI in hiring: Experimental evidence on how biased hiring algorithms and different debiasing methods affect the quality and diversity of applicants. Behavioral Science & Policy, 11(1), 44–54.

Irlenbusch, B., Rau, H. A., & Rilke, R. M. (2026). Human–AI evaluation and gender transparency: Application decisions in competitive hiring (ECONtribute Discussion Paper No. 398).

Ivanova-Stenzel, R. (2025). Delegating in the age of AI: Preferences for decision autonomy. Discussion Paper No. 558.

Ivanova-Stenzel, R., & Tolksdorf, M. (2024). Measuring preferences for algorithms—How willing are people to cede control to algorithms? Journal of Behavioral and Experimental Economics, 112, 102270.

Jabarian, B., & Henkel, L. (2026). Voice AI in firms: A natural field experiment on automated job interviews. Working Paper.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 624–635.

Jiang, W., Li, D., & Liu, C. (2025). Understanding dimensions of trust in AI through quantitative cognition: Implications for human-AI collaboration. PLoS One, 20(7), e0326558.

Johnson, T., & Obradovich, N. (2024). Measuring an artificial intelligence agent's trust in humans using machine incentives. Journal of Physics: Complexity, 5(1), 015003. https://doi.org/10.1088/2632-072X/ad1c69

Jung, M., & Seiter, M. (2021). Towards a better understanding on mitigating algorithm aversion in forecasting: An experimental study. Journal of Management Control, 32, 495–516.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. Proceedings of the 28th European Conference on Information Systems (ECIS).

Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. MIS Quarterly, 48(4), 1575–1590.

Kanazawa, K., Kawaguchi, D., Shigeoka, H., & Watanabe, Y. (2026). AI, skill, and productivity: The case of taxi drivers. Management Science, 72(2), 1376–1388.

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. Human Factors, 65(2), 337–359.

Kawaguchi, K. (2021). When will workers follow an algorithm? A field experiment with a retail business. Management Science, 67(3), 1670–1695.

Keppeler, F. (2024). No thanks, dear AI! Understanding the effects of disclosure and deployment of artificial intelligence in public sector recruitment. Journal of Public Administration Research and Theory, 34, 39–52.

Kim, H., Glaeser, E. L., Hillis, A., Kominers, S. D., & Luca, M. (2024). Decision authority and the returns to algorithms. Strategic Management Journal, 45(4), 619–648.

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. Computers in Human Behavior, 160, 108352.

Klockmann, V., von Schenk, A., & Villeval, M. C. (2022). Artificial intelligence, ethics, and intergenerational responsibility. Journal of Economic Behavior & Organization, 203, 284–317.

Klockmann, V., von Schenk, A., & Villeval, M. C. (2025). Artificial intelligence, distributional fairness, and pivotality. European Economic Review, 178, 105098.

Krakowski, S., Haftor, D., Luger, J., Pashkevich, N., & Raisch, S. (2026). Human-centered artificial intelligence: A field experiment. Management Science, 72(1), 57–72.

Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J. F., & Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. Nature, 646, 126–134. https://doi.org/10.1038/s41586-025-09505-x.

Lane, T., Pickard, H., & Walker, M. J. (2025). No silver lining: Consumer indifference between human and AI production, SSRN 5184807.

Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. Computers in Human Behavior, 123, 106878.

Lee, Y.-S., & Weber, R. A. (2025). Revealed privacy preferences: Are privacy choices rational? Management Science, 71(3), 2657–2677.

Lehr, S. A., Cipperman, M., & Banaji, M. R. (2026). Extreme Self-Preference in Language Models. Discussion Paper, Harvard University.

Leib, M., Köbis, N., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2024). Corrupted by algorithms? How AI-generated and human-written advice shape (dis)honesty. The Economic Journal, 134, 766–784.

Li, L., Wei, W., & Yao, Y. (2026). Social class and LLM adoption (Unpublished manuscript).

Liao, Q. V., & Sundar, S. S. (2022). Designing for responsible trust in AI systems: A communication perspective. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), 1257–1268.

Lin, T. (2022). Valuing intrinsic and instrumental preferences for privacy. Marketing Science, 41(4), 663–681.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103.

Luo, X., Qin, M. S., Fang, Z., & Qu, Z. (2021). Artificial intelligence coaches for sales agents: Caveats and solutions. Journal of Marketing, 85(2), 14–32.

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. Technological Forecasting and Social Change, 175, Article 121390. https://doi.org/10.1016/j.techfore.2021.121390

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734.

Mehrotra, S., Degachi, C., & Vereschak, O. (2024). A systematic review on fostering appropriate trust in human-AI interaction: Trends, opportunities and challenges. ACM Journal on Responsible Computing, 1(4), 1–45.

Opitz, T., Sliwka, D., Vogelsang, T., & Zimmermann, T. (2025). The algorithmic assignment of incentive schemes. Management Science, 71(2), 1546–1563.

Orbán, F., & Stefkovics, Á. (2025). Trust in artificial intelligence: A survey experiment to assess trust in algorithmic decision-making. AI & SOCIETY, 40, 4955–4969.

Pethig, F., & Kroenung, J. (2023). Biased humans, (un)biased algorithms? Journal of Business Ethics, 183, 637–652.

Pisanelli, E. (2022). Your resume is your gatekeeper: Automated resume screening as a strategy to reduce gender gaps in hiring. Economics Letters, 221, 110892.

Riedl, R. (2022). Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions. Electronic Markets, 32(4), 2021–2051.

Rostam-Afschar, D., Brüll, E., & Mäurer, S. (2025). Beliefs about bots: How employers plan for AI in white-collar work (Discussion Paper No. 25-057). ZEW Leibniz Centre for European Economic Research.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. Academy of Management Review, 23(3), 393–404.

Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. Science and Engineering Ethics, 26, 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. Journal of Decision Systems, 29(4), 260–278.

Strobel, C. (2025). Impact of process automation on performance. Journal of Behavioral and Experimental Economics, 117, 102377.

Sun, F., Li, N., Wang, K., & Goette, L. (2025). Large language models are overconfident and amplify human bias. arXiv. https://doi.org/10.48550/arXiv.2505.02151

Svirsky, D. (2022). Privacy and information avoidance: An experiment on data-sharing preferences. Journal of Legal Studies, 51(1), 63–92.

Tomaino, G., Wertenbroch, K., & Walters, D. J. (2023). Intransitivity of consumer preferences for privacy. Journal of Marketing Research, 60(3), 489-507.

Wang, H., Zhang, Y., & Lu, T. (2026). The power of disagreement: A field experiment to investigate human-algorithm collaboration in loan evaluations. Management Science, 72(1), 96–118.

Wang, S. Q., Adar, E., & Chen, Y. (2025). Privacy with information externalities and complexity. SSRN.

Xu, C., Dai, T., & Yan, X. (2026). Identity disclosure and anthropomorphism in voice chatbot design: A field experiment. Management Science, 72(1), 223–241.

Xu, J., Li, G., & Jiang, J. Y. (2025). AI self-preferencing in algorithmic hiring: Empirical evidence and insights. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8(3), 2757–2758.

Zhang, S., & Kuhn, P. (2024). Measuring bias in job recommender systems: Auditing the algorithms (IZA DP No. 17245). IZA Institute of Labor Economics.

Zhang, S., & Narayandas, D. (2026). Engaging customers with AI in online chats: Evidence from a randomized field experiment. Management Science, 72(1), 73–95.