Why People Reject Good Algorithms: Overcoming Algorithm Aversion in the Workplace

Executive summary

Two Wharton-led studies from the Wharton School[1] remain the clearest foundation for understanding algorithm aversion. The first showed that people become especially reluctant to use an algorithm after seeing it make a mistake, even when it still outperforms a human forecaster; the second showed that this aversion can be reduced if users are allowed to modify the algorithm's output, even slightly. In other words, the problem is not simply that people dislike algorithms. It is that they are less forgiving of visible algorithmic errors and more willing to accept imperfect outcomes when they retain a sense of agency. [2]

That is only half the story. Later research on algorithm appreciation found that people sometimes prefer algorithmic advice to human advice, especially for structured, impersonal, numeric tasks. More recent work from the MIT Sloan School of Management[3] reconciles the "aversion versus appreciation" debate: people tend to prefer AI when it is perceived as more capable than humans **and** the task does not require personalization; when either condition fails, people revert toward human judgment. This is why the same organization can see fast adoption in fraud detection and deep resistance in hiring, performance reviews, diagnosis, or other identity-laden decisions. [4]

From a behavioral-economics perspective, the core drivers are loss aversion, status quo bias, and overconfidence. Prospect theory predicts that losses loom larger than gains, which helps explain why one salient algorithmic miss can erase the memory of many correct outputs. Status quo bias makes incumbent human workflows feel safer and more legitimate than changed AI-assisted workflows. Overconfidence makes users believe their own case-by-case intuition is better than it really is, especially when they have domain familiarity or interpretability cues that increase the illusion of understanding. [5]

For product managers and UX researchers, the strongest implication is that adoption should not be optimized through raw acceptance of AI recommendations. The right objective is **appropriate reliance**: users should follow the model when it is likely to be right and override it when it is likely to be wrong. That requires bounded override mechanisms, calibrated confidence displays, layered explanations that reveal both rationale and limitations, and friction only on risky disagreements rather than everywhere. Success should be measured with metrics such as blind acceptance of incorrect recommendations, harmful overrides of correct recommendations, decision-quality lift, and audit-compliant override behavior. [6]

What the evidence says

The original Wharton result is crisp: people lose confidence in algorithmic forecasters more quickly than human forecasters after seeing the same mistake. That asymmetry is the empirical heart of algorithm aversion. The follow-up Wharton paper then showed that people are more likely to choose an imperfect algorithm if they can modify it, and that this preference appears to reflect a desire for **some** control rather than a need for large amounts of control. Even severe limits on permitted adjustment did not erase the effect. That finding is product-relevant because it suggests that a carefully designed override can increase adoption without surrendering the entire decision to ad hoc judgment. [7]

Manager-facing commentary in Harvard Business Review[8] helped diffuse this insight into practice by emphasizing a common user intuition: many people expect human judgment to learn and adapt more naturally than a formula. That intuition matters because it is often strongest where users feel that context, empathy, or exception handling matter most. It also matches the later MIT capability-personalization framework: perceived capability alone is not enough; users also ask whether the task calls for human recognition of uniqueness or special circumstances. [9]

The literature then moved from a one-sided "people dislike algorithms" story to a more conditional view. Algorithm appreciation research found that, in several numeric-estimation settings, people adhered more to advice when they believed it came from an algorithm than from a person. But the same paper also found that experienced professionals relied **less** on algorithmic advice than laypeople, and that this reduced their accuracy. This matters in workplaces because resistant users are often not novices. They are experienced operators whose self-confidence, identity, and familiarity with the task can make them selectively skeptical of model advice. [10]

A useful way to frame the behavioral economics behind these findings is to separate three mechanisms. First, **loss aversion** makes a visible AI mistake feel disproportionately costly relative to the many quiet times the system was right; a recent Management Science paper also found that loss framing can eliminate an initial preference for human over superior AI assistance. Second, **status quo bias** makes "keeping the human process" the psychologically and politically easier option; managerial studies show that value, tradition, and image barriers are significantly associated with algorithm aversion. Third, **overconfidence** leads decision makers to overestimate their own performance, overplace themselves relative to others, or express excessive precision; in AI-assisted settings, that shows up as second-guessing models, overvaluing one's own local intuition, or misreading interpretability as true mastery. [11]

Why algorithm errors feel worse than human bias

The literature implies a deep asymmetry between **error perception** and **bias perception**. Algorithmic errors are salient, discrete, attributable, and easy to replay. Human bias is often chronic, diffuse, socially normalized, and partially hidden inside judgment calls,

memory, or inconsistent weighting of evidence. So users often punish the algorithm for a visible miss while tolerating a much larger background rate of human inconsistency. This is exactly why classic work comparing actuarial and clinical judgment found that properly developed statistical methods often perform as well as or better than expert judgment, yet experts continue to trust unvalidated personal impressions. [12]

Renier and colleagues sharpen this point: when an error is made by an algorithm rather than a human, reactions are harsher, negative feelings are stronger, and behavioral intentions tied to action are stronger, while forgiveness and accountability judgments are weaker. Put differently, people do not merely trust the erring algorithm less. They also react to it as a qualitatively different kind of failure. At the same time, human judgment continues to benefit from social excuses: "the case was complex," "the manager used discretion," or "the clinician saw something the model missed." [13]

The table below synthesizes the main asymmetries that matter in product design. It draws on the Wharton studies, research on reactions to erring algorithms, classic comparisons of actuarial versus clinical judgment, and more recent work on human-AI oversight. [14]

Dimension	How users often perceive algorithm errors	How users often perceive human errors or bias	Product implication
Salience	A single wrong output is vivid and memorable	Human inconsistency is spread out across many cases	Show aggregate performance and known failure modes, not just isolated outputs
Attribution	"The system failed" feels concrete and attributable	"The person used judgment" softens blame	Add rationale, confidence, and context so failures are interpretable rather than mysterious
Forgiveness	Lower tolerance after equal mistakes	Greater willingness to excuse or contextualize mistakes	Design recovery paths that preserve trust after recoverable errors
Learning model	Users doubt the model after observed error	Users assume humans can learn and adapt	Communicate model updates, retraining logic, and domain boundaries clearly
Hidden bias	Algorithmic bias is feared as systemic	Human bias is often treated as ordinary discretion	Audit both human overrides and model outputs; do not assume human oversight is neutral
Oversight assumption	"A human can always catch bad	Human review is assumed to be a	Instrument when reviewers actually detect errors

Dimension	How users often perceive algorithm errors	How users often perceive human errors or bias	Product implication
	AI”	safety net	instead of assuming they do

A crucial further nuance is that human oversight is not automatically corrective. Research on algorithm-in-the-loop decision making found that human behavior often improved accuracy but failed to satisfy stronger reliability and fairness criteria. Other studies show that people can inherit AI biases and continue reproducing them even after the AI is removed from the immediate task. This means a product team should not treat “human in the loop” as sufficient governance; the loop itself has to be designed, measured, and audited. [15]

Design patterns for workplace products

The most useful design guidance comes from Microsoft Research[16]’s human-AI interaction guidelines and the National Institute of Standards and Technology[17] explainability framework, combined with the workplace and decision-support studies reviewed above. The central lesson is that no single feature—override, explanation, or confidence—solves trust by itself. These features work only when they are matched to user cognition, task structure, and error costs. [18]

The intervention comparison below synthesizes the most defensible patterns for workplace products. It translates the evidence into choices product teams can actually ship. [19]

Pattern	Why it helps	Best fit	Main trade-off	Implementation guidance
Constrained override	Restores user agency and reduces aversion	Forecasts, rankings, allocation, prioritization	Can improve acceptance while reducing pure model performance if edits are too free	Start with bounded edits such as adjacent-rank swaps, narrow score bands, or ±5–10% adjustments; log magnitude and direction of edits
Full override with reason code	Preserves discretion where stakes or accountability require it	Hiring, performance review, clinical or financial exceptions	Too little friction invites ego-driven or status-quo overrides; too much friction creates	Require concise structured reasons only on material disagreements or high-impact cases; add

Pattern	Why it helps	Best fit	Main trade-off	Implementation guidance
			resentment	second review for large deviations
Layered explanation	Helps users understand why and when the system acted	Medium- and high-stakes decisions	Dense explanations can increase blind trust or cognitive burden	Use three layers: short rationale inline, drill-down "why this case," and separate "when not to trust this output"
Confidence bands plus action guidance	Supports case-by-case trust calibration	Any task with calibrated probability estimates	Raw numeric confidence is hard to interpret and can backfire when miscalibrated	Prefer coarse bands such as High / Medium / Low paired with clear actions; expose raw percentages only after validating comprehension
Manual-first or hypothesis-first workflow	Reduces anchoring on AI and keeps users mentally engaged	High-personalization, high-risk, or ambiguous tasks	Adds time and effort	Elicit the user's initial view before showing the recommendation when critical thinking matters more than speed
Cognitive-forcing prompts	Reduces overreliance on wrong AI recommendations	Low-confidence, high-impact, or disagreement cases	Most effective designs often feel less pleasant and slower	Trigger only on risky cases; use checklists, "rule out alternatives," or short verification prompts rather than universal friction

User decision flow with overrides

```
flowchart TD
    A[Eligible case arrives] --> B{Known weak region,\nhigh personalization,\nor out-of-distribution signal?}
    B -- Yes --> C[Manual-first judgment]
    B -- No --> D[Model scores case]
```

```
    D --> E{Confidence and knowledge limits}
    E -- Low / uncertain --> C
    E -- Medium / high --> F[Show recommendation\n+ confidence band\n+ short
rationale\n+ known limits]
    C --> G[Capture user's initial view]
    G --> F
    F --> H{User action}
    H -- Accept --> I[Finalize decision]
    H -- Adjust within bounds --> J[Log bounded override]
    H -- Full override --> K{High impact or\nhigh-confidence disagreement?}
    J --> I
    K -- No --> L[Require structured reason]
    K -- Yes --> M[Require structured reason\n+ second review or checklist]
    L --> I
    M --> I
    I --> N[Audit trail and feedback loop:\naccuracy, override
quality,\ncalibration, compliance]
```

This flow reflects several consistent findings: ask for user input before the AI on cases where anchoring risk is costly; make the system's confidence and rationale visible; permit some control because agency improves acceptance; escalate only risky disagreements; and instrument the feedback loop so the organization can tell whether overrides improved or degraded outcomes. [20]

Three concrete feature recommendations follow from the literature.

First, **override mechanisms should be graduated, not binary**. A slight-override design is often better than a blunt "accept/reject" control for structured outputs because it captures the agency benefit documented by Wharton while limiting performance damage from unrestricted intuition. But on high-impact decisions, unrestricted overrides may still be necessary for governance reasons; in that case, reason codes, visible policy thresholds, and post hoc audits become essential. The design principle is not "remove human discretion." It is "make discretion legible, reviewable, and proportionate to risk." [21]

Second, **explainability should be layered and limitation-aware**. The evidence does not support a blanket assumption that more explanation always improves judgment. IBM's decision-support experiments found that confidence information helped calibrate trust, while local feature-level explanations did not reliably improve outcomes. Harvard's cognitive-forcing work found that designs that most reduced overreliance were often less liked. And workplace field evidence from MIT Sloan suggests that interpretability can sometimes increase second-guessing and reduce acceptance under uncertainty. That argues for an explanation stack that is concise by default, deeper on demand, and explicit about failure modes or knowledge limits—not a wall of SHAP bars shown on every case. [22]

Third, **confidence scores should be treated as interaction design, not just model output**. Well-calibrated confidence can help users know when to rely and when to override. But confidence signals are often miscalibrated, users struggle to interpret them,

and overconfident wrong predictions can actively worsen trust and decision quality. A practical rule is to expose only those confidence signals that have passed both model-level calibration checks and user-comprehension testing. When confidence is poor or unstable, it is better to hide the number, abstain, or communicate uncertainty qualitatively than to display a misleading precision. [23]

Trust calibration and measurement

Appropriate reliance is not the same as maximum trust. Classic trust-in-automation work defines the target as correspondence between trust and actual trustworthiness. In product terms, that means the right user should trust the right model output by the right amount at the right time. A system can have high adoption and still be badly designed if users are accepting the wrong recommendations or overriding the right ones. [24]

Two practical trust-calibration models are especially useful.

One is a **correctness-likelihood model**: on a given case, rely on AI when its calibrated likelihood of being correct exceeds the user's own expected correctness by a task-dependent threshold. The threshold should rise with the cost of an AI error and fall when human review is expensive but low value. This extends recent work arguing that case-level trust should consider the relative correctness likelihood of both the AI and the human, not just whether AI confidence exceeds a fixed threshold. [25]

$$\text{Use AI on case } i \quad \text{if} \quad p_{AI,i} - p_{H,i} > \tau_i$$

Here, $p_{AI,i}$ is the calibrated probability that the model is correct on case i, $p_{H,i}$ is the estimated probability that the user or reviewer is correct on that case type, and τ_i is a risk threshold that grows when false acceptance is costly.

The second is a **reliance-gap model** from recent decision-theoretic work. Let reliance be the probability that the final actor chooses the AI recommendation when the AI and human disagree. Then behavioral overreliance exists when observed reliance exceeds the rational, payoff-maximizing reliance level; underreliance exists when it falls below that benchmark. This is a strong PM/UX lens because it moves evaluation away from attitudes and toward decision quality under disagreement. [26]

$$\gamma = Pr(a = y^{AI} \mid y^{AI} \neq y^{H})$$

$$\text{Overreliance if } \gamma^b > \gamma^r, \qquad \text{Underreliance if } \gamma^b < \gamma^r$$

$$\Delta = R - R_{\emptyset}$$

In this notation, γ^b is observed behavioral reliance, γ^r is rational reliance, and Δ is the maximum achievable complementarity lift above the better of human-alone or AI-alone decision making.

Finally, there is **confidence calibration health**. Teams should monitor Expected Calibration Error as a model-level health metric and review reliability diagrams by user segment and case type. A common operationalization is:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$$

where each B_m is a confidence bin. The point is not that PMs must become calibration theorists. The point is that a confidence number should not appear in the interface unless modelers can show it is informative at the level users will actually see and act on. [27]

The KPI map below operationalizes these ideas for adoption and compliance. It synthesizes overreliance metrics, weight-of-advice measures, and trust-calibration frameworks from the human-AI literature. [28]

KPI	Definition	What it diagnoses	Desired direction
Eligible-case usage rate	AI-assisted sessions / eligible sessions	Adoption and workflow penetration	Up, but not as a sole success metric
Appropriate reliance rate	Accept correct AI + reject incorrect AI, divided by disagreement cases	Net calibration of human-AI teamwork	Up
Blind acceptance rate	Accepted incorrect AI recommendations / incorrect AI recommendations shown	Overreliance	Down
Harmful override rate	Rejected correct AI recommendations / correct AI recommendations in disagreement cases	Underreliance or ego-driven discretion	Down
Weight of advice	Degree to which final judgment moves toward AI advice	Strength of AI influence in forecast-style tasks	Context-dependent; compare to rational benchmark
Decision-quality lift	Final decision quality minus better of human-alone or AI-alone baseline	Whether the team is actually complementary	Up
Override justification completeness	Overrides with policy-compliant reason code / all material overrides	Compliance and auditability	Up
Escalation	Cases meeting escalation	High-risk	Up

KPI	Definition	What it diagnoses	Desired direction
adherence	rules that actually received second review	governance quality	
Time to final decision	Median or p90 completion time	Friction cost of trust-calibration design	Stable or slightly down, unless deliberate review is added
Confidence comprehension score	Survey or in-product quiz on confidence meaning	Whether users can interpret uncertainty correctly	Up

For experimentation, the strongest A/B tests are not generic "show explanation vs no explanation" tests. They are behaviorally targeted tests around **sequence**, **agency**, **confidence presentation**, and **risky disagreement handling**. Pre-register the primary metric and stratify or block by user expertise, task familiarity, personalization level, and case confidence. Those moderators are repeatedly shown to change reliance behavior. [29]

For binary primary outcomes such as adoption, blind acceptance, or harmful overrides, a practical approximation for user-level randomization is:

$$n_{\text{per arm}} \approx \frac{\left(z_{1-\alpha/2}\sqrt{2\bar{p}(1-\bar{p})} + z_{1-\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}\right)^2}{(p_2-p_1)^2}$$

With $\alpha = 0.05$ and power $= 0.80$, that yields roughly these per-arm sample sizes: about **1,530** users to detect a 5-point increase in adoption from 55% to 60%; about **1,030** to detect a 6-point increase in appropriate reliance from 35% to 41%; and about **1,450** to detect a 4-point reduction in harmful blind acceptance from 20% to 16%. For three-arm comparisons with pairwise testing, a rough planning number for a 5-point adoption effect is about **2,050** users per arm. These are planning estimates and should be increased for attrition, repeated-measures clustering, or strong segment-level analyses.

The experimentation matrix below translates the literature into concrete product tests.

Experiment	Variants	Primary metric	Key secondary metrics	Why this test matters
Sequence test	AI-first vs manual-first vs manual-first + confidence band	Appropriate reliance rate	Time to decision, harmful override, user confidence	Tests anchoring reduction and whether asking for initial judgment improves calibration

Experiment	Variants	Primary metric	Key secondary metrics	Why this test matters
Override test	Full free override vs bounded adjustment + full override with reason code	Harmful override rate	Adoption, satisfaction, audit completeness	Tests the Wharton "some control" effect against governance needs
Confidence UI test	No confidence vs raw percentage vs three-band confidence with action guidance	Blind acceptance rate	Appropriate reliance, comprehension score, adoption	Tests whether uncertainty helps only when users can understand it
Explanation test	Always-on local explanation vs on-demand layered explanation vs layered explanation + known limits	Decision-quality lift or appropriate reliance	Trust, response time, satisfaction	Tests the difference between informative explanation and merely persuasive explanation
Risk escalation test	No extra review vs reason code only vs reason code + second review on risky disagreements	Audit-compliant override rate	Blind acceptance, harmful override, turnaround time	Tests whether selective friction improves safety without depressing normal-case adoption

A final practical guideline: use **one** primary metric per experiment. For adoption experiments, that should usually be usage rate or accept rate on eligible cases. For trust-calibration experiments, it should usually be appropriate reliance or blind acceptance. For governance experiments, it should usually be override quality or escalation adherence. Secondary metrics are necessary guardrails, but they should not be allowed to blur the causal readout of the test.

References

- *Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err* — Journal of Experimental Psychology: General, 2015. Foundational Wharton-led paper establishing stronger confidence loss after equal algorithm versus human mistakes. [30]

- *Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them* — Management Science, 2018. Wharton-led evidence that even slight user control can materially reduce aversion. [31]
- *Algorithm Appreciation: People Prefer Algorithmic to Human Judgment* — Organizational Behavior and Human Decision Processes, 2019. Shows appreciation can occur, especially in structured estimation contexts, while experts may rely less than laypeople. [32]
- *AI Aversion or Appreciation? A Capability–Personalization Framework and a Meta-Analytic Review* — Psychological Bulletin, summarized by MIT News and MIT Sloan, 2025. Reconciles aversion and appreciation through perceived capability and perceived need for personalization. [33]
- *Trust in Automation: Designing for Appropriate Reliance* — Human Factors, 2004. Foundational framework defining trust calibration as appropriate reliance rather than maximal trust. [34]
- *Guidelines for Human-AI Interaction* — CHI, 2019. Practical design guidance on dismissal, correction, explanation, feedback, and controls. [35]
- *Four Principles of Explainable Artificial Intelligence* — NIST IR 8312, 2021. Defines explanation, meaningfulness, explanation accuracy, and knowledge limits. [36]
- *Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making* — FAT*, 2020. Finds confidence scores can help calibrate trust, while local explanations do not reliably improve decision outcomes. [37]
- *Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance* — HCOMP, 2019. Introduces the idea that users need a mental model of the AI's error boundary. [38]
- *To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making* — Proceedings of the ACM on Human-Computer Interaction, 2021. Shows effective anti-overreliance interventions often create usability and preference costs. [39]
- *The Principles and Limits of Algorithm-in-the-Loop Decision Making* — Proceedings of the ACM on Human-Computer Interaction, 2019. Shows that adding humans to the loop does not automatically deliver reliable or fair oversight. [40]
- *Humans Inherit Artificial Intelligence Biases* — Scientific Reports, 2023. Demonstrates that users can internalize biased AI recommendations and reproduce them later without AI assistance. [41]
- *To Err Is Human, Not Algorithmic* — Computers in Human Behavior, 2021. Shows harsher affective and behavioral reactions to erring algorithms than to erring humans. [13]
- *The Trouble With Overconfidence* — Psychological Review, 2008. Clear framework separating overestimation, overplacement, and overprecision. [42]

- *Prospect Theory: An Analysis of Decision Under Risk* — Econometrica, 1979. Foundational account of loss aversion and asymmetric valuation of losses versus gains. [43]
- *Status Quo Bias in Decision Making* — Journal of Risk and Uncertainty, 1988. Classic evidence that people disproportionately stick with existing options and defaults. [44]
- *Humans' Use of AI Assistance: The Effect of Loss Aversion on Willingness to Delegate Decisions* — Management Science, 2025 online / 2026 issue. Finds loss framing can remove an initial human-over-AI delegation bias. [45]
- *A Decision Theoretic Framework for Measuring AI Reliance* — ACM manuscript/preprint, 2024. Formalizes appropriate reliance, overreliance, underreliance, and complementarity lift. [46]
- *Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making* — ACM manuscript/preprint, 2023. Argues that trust calibration should consider both human and AI case-level correctness likelihood. [47]
- *Understanding the Effects of Miscalibrated AI Confidence on User Trust, Reliance, and Decision Efficacy* — arXiv manuscript, 2025 version. Shows users struggle to detect miscalibrated confidence and can over- or under-rely accordingly. [48]
- *On Calibration of Modern Neural Networks* — ICML, 2017. Foundational calibration paper and standard source for ECE and temperature scaling. [49]
- *What Drives Managers Towards Algorithm Aversion and How to Overcome It?* — Technological Forecasting and Social Change, 2023. Finds value, tradition, and image barriers correlate with managerial aversion, mitigated partly by technology readiness. [50]
- *Here's Why People Trust Human Judgment Over Algorithms* — Harvard Business Review, 2015. Early managerial translation of the algorithm-aversion finding for executives. [51]

---

[1] [5] [43] https://web.mit.edu/curhan/www/docs/Articles/15341_Readings/Behavioral_Decision_Theory/Kahneman_Tversky_1979_Prospect_theory.pdf

https://web.mit.edu/curhan/www/docs/Articles/15341_Readings/Behavioral_Decision_Theory/Kahneman_Tversky_1979_Prospect_theory.pdf

[2] [7] [12] [14] [30] https://marketing.wharton.upenn.edu/wp-content/uploads/2016/10/Dietvorst-Simmons-Massey-2014.pdf

https://marketing.wharton.upenn.edu/wp-content/uploads/2016/10/Dietvorst-Simmons-Massey-2014.pdf

[3] [41] https://www.nature.com/articles/s41598-023-42384-8

https://www.nature.com/articles/s41598-023-42384-8

[4] [10] [32] https://www.jennlogg.com/uploads/2/8/9/2/2892148/algorithm_appreciation__logg_minson_moore_2019_.pdf

https://www.jennlogg.com/uploads/2/8/9/2/2892148/algorithm_appreciation__logg_minson_moore_2019_.pdf

[6] [17] [28] https://www.microsoft.com/en-us/research/wp-content/uploads/2022/06/Aether-Overreliance-on-AI-Review-Final-6.21.22.pdf

https://www.microsoft.com/en-us/research/wp-content/uploads/2022/06/Aether-Overreliance-on-AI-Review-Final-6.21.22.pdf

[8] [26] [46] https://mucollective.northwestern.edu/files/2024-AI_reliance.pdf

https://mucollective.northwestern.edu/files/2024-AI_reliance.pdf

[9] [51] https://hbr.org/2015/02/heres-why-people-trust-human-judgment-over-algorithms

https://hbr.org/2015/02/heres-why-people-trust-human-judgment-over-algorithms

[11] [45] https://pubsonline.informs.org/doi/10.1287/mnsc.2024.05585

https://pubsonline.informs.org/doi/10.1287/mnsc.2024.05585

[13] https://www.sciencedirect.com/science/article/pii/S0747563221002028

https://www.sciencedirect.com/science/article/pii/S0747563221002028

[15] [40] https://www.benzevgreen.com/wp-content/uploads/2019/09/19-cscw.pdf

https://www.benzevgreen.com/wp-content/uploads/2019/09/19-cscw.pdf

[16] [22] [23] [37] https://qveraliao.com/fat2020.pdf

https://qveraliao.com/fat2020.pdf

[18] [35] https://www.microsoft.com/en-us/research/wp-content/uploads/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf

https://www.microsoft.com/en-us/research/wp-content/uploads/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf

[19] [21] [31] https://faculty.wharton.upenn.edu/wp-content/uploads/2016/08/Dietvorst-Simmons-Massey-2018.pdf

https://faculty.wharton.upenn.edu/wp-content/uploads/2016/08/Dietvorst-Simmons-Massey-2018.pdf

[20] https://arxiv.org/pdf/1802.07810

https://arxiv.org/pdf/1802.07810

[24] [34] https://journals.sagepub.com/doi/10.1518/hfes.46.1.50_30392

https://journals.sagepub.com/doi/10.1518/hfes.46.1.50_30392

[25] [47] https://arxiv.org/pdf/2301.05809

https://arxiv.org/pdf/2301.05809

[27] [49] https://proceedings.mlr.press/v70/guo17a/guo17a.pdf

https://proceedings.mlr.press/v70/guo17a/guo17a.pdf

[29] https://mitsloan.mit.edu/press/why-our-minds-sometimes-say-no-even-when-ai-right

https://mitsloan.mit.edu/press/why-our-minds-sometimes-say-no-even-when-ai-right

[33] https://news.mit.edu/2025/how-we-really-judge-ai-0610

https://news.mit.edu/2025/how-we-really-judge-ai-0610

[36] https://nvlpubs.nist.gov/nistpubs/ir/2021/nist.ir.8312.pdf

https://nvlpubs.nist.gov/nistpubs/ir/2021/nist.ir.8312.pdf

[38] https://erichorvitz.com/gbansal-hcomp19.pdf

https://erichorvitz.com/gbansal-hcomp19.pdf

[39] https://www.eecs.harvard.edu/~kgajos/papers/2021/bucinca21trust.pdf

https://www.eecs.harvard.edu/~kgajos/papers/2021/bucinca21trust.pdf

[42] https://healy.econ.ohio-state.edu/papers/Moore_Healy-TroubleWithOverconfidence.pdf

https://healy.econ.ohio-state.edu/papers/Moore_Healy-TroubleWithOverconfidence.pdf

[44] https://link.springer.com/article/10.1007/BF00055564

https://link.springer.com/article/10.1007/BF00055564

[48] https://arxiv.org/html/2402.07632v4

https://arxiv.org/html/2402.07632v4

[50] https://www.sciencedirect.com/science/article/pii/S0040162523003268

https://www.sciencedirect.com/science/article/pii/S0040162523003268