When Should People Trust an AI's Recommendation?
A Pre-Registered Design for Studying Appropriate Reliance

Human–AI Trust Lab

September 7, 2026
Draft manuscript — no human-subject data collected

Evidence status. This draft documents a completed experimental design, engine, and analysis pipeline, validated only against a labelled participant *simulator* (Section 10). No human-subject data have been collected. Every number in Section 8 is generated from `results/*.json` by `research/scripts/05_paper_export.py` and is explicitly a simulation result, not a finding about human behaviour.

Abstract

Automated decision aids are increasingly deployed alongside human judgment, yet the same literature reports both *automation bias* (over-following imperfect algorithms) and *algorithm aversion* (under-using algorithms that outperform people), often in similar task settings. We present a pre-registered, repeated-decision experimental design and a complete, tested engineering pipeline for studying when each failure mode occurs, manipulating AI accuracy and confidence-display calibration between subjects and AI correctness, confidence level and task difficulty within subjects, across two task families (perceptual dot comparison and probabilistic series forecasting). We define reliance outcomes that separate raw acceptance from *appropriate* reliance (RAIR, RSR), specify a primary generalized-estimating-equations analysis on disagreement trials, and validate every estimator against an archetype-based participant simulator with known ground truth before any human data are collected. At the time of writing no human-subject data exist; we report validation results only, and the manuscript is structured so that every quantitative claim is regenerated automatically from a versioned, hash-checked dataset once real data are available.

1 Introduction

Human decision-makers increasingly receive recommendations from AI systems in domains ranging from content moderation to medical triage. Two well-replicated but seemingly contradictory findings describe how people respond: Dietvorst et al. [2015] show that people abandon even well-performing algorithms after seeing them make a single visible error ("algorithm aversion"), while a substantial applied literature documents the opposite failure mode, *automation bias* [Parasuraman and Riley, 1997], in which people follow an automated recommendation past the point where independent judgment would have caught its error. More recent work finds contexts in which people instead over-trust algorithmic judgment relative to human judgment ("algorithm appreciation", Logg et al., 2019).

These results are not necessarily inconsistent: they may reflect different points on the same underlying surface, indexed by AI accuracy, how confidence is communicated, whether an explanation is given, and whether the person has had a chance to learn the AI's reliability. Our goal is to trace

that surface directly, in one coherent experimental platform, rather than adding another single-condition data point to a fragmented literature. We designate two purpose-built task families, a factorial design chosen to be analyzable at reasonable sample sizes (Section 4), and — because our own trial-level pipeline is exactly the kind of system that can silently encode the wrong estimand — a battery of simulation-based validations that must pass before any claim about human behaviour is entertained (Section 10).

This manuscript documents the design, the complete software pipeline (experiment engine, statistical analysis, and validation), and reports only what currently exists: a validated pipeline with zero human-subject observations. We are explicit throughout about what is established (the design is implementable and the estimators are unbiased on simulated data with known truth) and what is not yet established (anything about actual human reliance).

2 Related Work

Automation bias and algorithm aversion. Parasuraman and Riley [1997] catalogue over-reliance on automation across domains; Dietvorst et al. [2015] show a single observed error is enough to reduce reliance on an algorithm below reliance on an equally error-prone human, an effect reversed when the human retains override authority. Logg et al. [2019] instead find algorithm appreciation across judgment tasks lacking strong domain expertise, suggesting task and framing moderate the direction of the bias.

Confidence and explanation. Zhang et al. [2020] manipulate stated model confidence and local explanations directly and find confidence displays increase reliance and can improve trust calibration, while explanations alone do not reliably improve appropriate reliance — motivating our H2/H3 distinction between reliance and *appropriate* reliance. Lai and Tan [2019] similarly find explanation format changes reliance without necessarily improving joint accuracy. Buçinca et al. [2021] show that forcing a delay or an independent judgment before revealing AI advice (a "cognitive forcing function") reduces overreliance, which we treat as a candidate future manipulation rather than a factor in the present design.

Team performance and complementarity. Bansal et al. [2021] formalize complementary team performance as the gap between team accuracy and the better of the two individual accuracies, and show explanations can move this gap either way depending on whether they increase appropriate or blanket reliance — the same construct we use for H7. Yin et al. [2019] study how stated model accuracy shapes trust and reliance, closely related to our AI-accuracy manipulation, but do not manipulate confidence display or track within-session learning, which are two of this design's contributions.

A shared outcome vocabulary. Schemmer et al. [2023] propose relative AI reliance and relative self-reliance as measures that isolate appropriate switching/staying behaviour from raw acceptance; we adopt these definitions (Section 6) so that results here are comparable to that and future work using the same constructs.

Novelty boundary. No single prior study we are aware of crosses AI accuracy with confidence-display calibration (calibrated vs. deliberately miscalibrated, rather than merely present vs. absent) while tracking within-session learning across two independent task families. That combination —

not any one factor alone — is this design's contribution; the individual factors are each well studied in isolation, as above.

3 Research Questions and Hypotheses

RQ. When should people rely on an AI recommendation, and how do AI accuracy, confidence display, explanation, feedback and task difficulty shape appropriate reliance, over-reliance, under-reliance, calibration and learning?

H1 Higher AI accuracy increases reliance.

H2 AI confidence increases reliance even after controlling for actual correctness (tested separately in the calibrated arm, where this is normatively justified, and the miscalibrated arm, where it is a pure confidence heuristic).

H3 Explanations increase reliance but do not necessarily increase *appropriate* reliance (experiment v2).

H4 Confidently-wrong AI recommendations produce stronger automation-bias failures (over-reliance) than uncertain-wrong recommendations.

H5 Immediate feedback improves long-run calibration relative to no feedback (experiment v2).

H6 Participants learn AI reliability heterogeneously (exploratory).

H7 Excess reliance reduces team performance when the AI is imperfect (the 0.70-accuracy arm), via lower complementarity as over-reliance rises.

These are treated as the starting, falsifiable hypotheses this design is built to test, not as foregone conclusions; Section 8 reports estimates only from simulated data with known properties, and none of the above is asserted about human participants in this draft.

4 Experimental Design

Experiment v1 — reliance & confidence (primary)

A 2×3 between-subject design crossing AI accuracy (0.70, 0.90) with confidence display (none; numeric and calibrated, i.e. $P(\text{correct} \mid \text{shown } c) = c$ by construction; numeric and miscalibrated, drawn from the correct-conditional distribution regardless of correctness). Within each session, 60 trials split into two 30-trial blocks — one per task family, order counterbalanced by participant seed — with difficulty (easy/medium/hard) balanced exactly within each block. AI accuracy is realized exactly per block (not merely in expectation): with 30 trials and accuracy A, exactly $\text{round}((1 - A) \times 30)$ recommendations are wrong, split evenly across difficulty levels.

Experiment v2 — explanation & feedback

A 3×2 between-subject design (explanation: none, short rationale, uncertainty-aware rationale; feedback: immediate, absent) at a fixed AI accuracy of 0.80 and calibrated numeric confidence, sharing the same task families, trial structure and outcome definitions.

Task families

Dot comparison: two panels of dots at a controlled count ratio (1.50 / 1.20 / 1.08 for easy/medium/hard), ground truth is simply which panel has more. *Series forecast*: 24 observed points of a linear-trend-plus-noise series; the participant predicts whether the next value will be above or below the last observed value. Trend magnitude is chosen so the Bayes-optimal accuracy is exactly $\Phi(z)$ for z corresponding to 0.90/0.75/0.60 at the three difficulty levels, where Φ is the standard normal CDF. Neither task requires professional expertise.

Randomization and reproducibility

Every session is fully determined by a (seed, condition, design) triple via a seeded sfc32 generator with per-purpose forked streams, so the exact trial sequence, AI recommendations and confidence values can be reconstructed offline from three short strings — used by both the analysis package and the debrief screen. Condition assignment uses covariate-free minimisation against live cell counts when a persistent store is available, and falls back to seed-uniform assignment otherwise (e.g. the offline simulator).

5 Data

Every dataset used by this pipeline carries a manifest recording its `dataset_kind` (`synthetic`, `pilot`, or `real`), a SHA-256 of the trial-level CSV, the git commit of the code that produced it, and (for synthetic data) the archetype mixture and any parameter overrides used. Analysis code refuses to run if the checksum does not match the manifest. At the time of writing the only datasets in the repository are synthetic (`data/synthetic/*`); no `pilot` or `real` dataset exists. The trial-level schema (32 columns: identifiers, condition, stimulus parameters, AI recommendation and confidence, and the participant's initial and final choices, confidences, and timestamps) is documented in `docs/DATA_CODEBOOK.md` and shared verbatim between the browser engine, the server persistence layer, and the Python analysis package.

6 Measures

Let y_t be ground truth, a_t the AI's recommendation, h_t the participant's initial choice, and f_t the final choice on trial t. Initial/final/AI accuracy are the corresponding means of $\mathbf{1}[\cdot = y_t]$; team accuracy is final accuracy; complementarity is final accuracy minus the larger of initial and AI accuracy [Bansal et al., 2021].

On *disagreement* trials ($h_t \neq a_t$), we follow Schemmer et al. [2023]: relative AI reliance RAIR $= P(f_t = a_t \mid h_t \neq a_t, a_t = y_t)$ and relative self-reliance RSR $= P(f_t = h_t \mid h_t \neq a_t, a_t \neq y_t)$; over-reliance is $1 - \text{RSR}$ and under-reliance is $1 - \text{RAIR}$. Confidence calibration uses the Brier score [Brier, 1950] and expected calibration error over five equal-width bins. Post-error trust adjustment is the within-participant difference in acceptance on disagreement trials following an AI error versus following an AI success (immediate-feedback arms only). Full formal definitions, including the derivation of the AI's exactly-realized accuracy and analytically calibrated confidence distribution, are in `docs/METHODOLOGY.md`.

7 Statistical Methods

The primary specification is a trial-level logistic generalized estimating equation [Liang and Zeger, 1986] with an exchangeable working correlation and participant-clustered robust standard errors, fit on disagreement trials:

$$\begin{aligned}\operatorname{logit} P(\text{accept}_t = 1) = \beta_0 + \beta_1 \text{acc90} + \beta_2 \text{cal} + \beta_3 \text{mis} + \beta_4 \text{aiCorrect}_t \\ + \beta_5 \tfrac{\text{shownConf}_t - 75}{10} + \gamma_{\text{difficulty}} + \gamma_{\text{family}} + \beta_6 z(\text{trial}_t).\end{aligned}$$

Every arm-specific confidence model additionally includes the accuracy-arm indicator, because displayed confidence is mechanically higher in the higher-accuracy arm by construction; omitting it confounds the confidence effect with the accuracy effect (Section 10 reports this as a validation finding, not a human result). H7 uses OLS with HC3 standard errors of complementarity on the participant-level over-reliance rate ($1 - \text{RSR}$), by accuracy arm. The confirmatory family {H1, H2-calibrated, H2-miscalibrated, H4, H7} is corrected with Holm's step-down procedure [Holm, 1979] at $\alpha = 0.05$; H3 and H5 belong to a separate design (experiment v2) and are not pooled into this family. Robustness checks include an independence working-correlation refit, a linear-probability model with participant fixed effects, exclusion-sensitivity analyses, and (sample size permitting) a random-intercept logistic model. All specifications, including the full robustness battery, are pre-specified in `analysis/README.md` before any human data are examined.

8 Results

Evidence status. This draft documents a completed experimental design, engine, and analysis pipeline, validated only against a labelled participant *simulator* (Section 10). No human-subject data have been collected. Every number in Section 8 is generated from `results/*.json` by `research/scripts/05_paper_export.py` and is explicitly a simulation result, not a finding about human behaviour.

Table 1: Descriptive outcomes by condition (SYNTHETIC — simulator validation, see evidence notice). Dataset sim-v1-001, sha 1507f368a1, code 57cb0777ac.

Cell	n	Initial acc.	AI acc.	Final acc.	Accept.	RAIR/RSR
AI 70%, none	40	76% [74%, 78%]	70% [70%, 70%]	77% [75%, 79%]	80% [77%, 84%]	71% / 57%
AI 70%, calibrated	40	77% [75%, 79%]	70% [70%, 70%]	79% [77%, 81%]	78% [75%, 81%]	71% / 66%
AI 70%, miscalibrated	40	75% [73%, 76%]	70% [70%, 70%]	78% [76%, 79%]	81% [78%, 84%]	77% / 59%
AI 90%, none	40	75% [73%, 77%]	90% [90%, 90%]	88% [86%, 90%]	90% [87%, 93%]	81% / 49%
AI 90%, calibrated	40	76% [75%, 78%]	90% [90%, 90%]	88% [86%, 90%]	91% [88%, 94%]	85% / 50%
AI 90%, miscalibrated	40	74% [72%, 76%]	90% [90%, 90%]	88% [86%, 90%]	91% [88%, 94%]	83% / 48%

9 Robustness

10 Estimator Validation

Because the trial-level engine, the persistence layer, and the analysis package are all custom software, we validate the full pipeline — not just the statistical test — against an archetype-based participant simulator (rational, over-truster, skeptic, confidence-sensitive, explanation-sensitive, noisy, learner, non-learner) that drives the actual experiment engine used by participants, not a simplified surrogate. Scenarios with a hard-coded parameter (e.g. confidence weight fixed to zero)

Table 2: Confirmatory family, Holm-corrected at $\alpha = 0.05$ (SYNTHETIC — see evidence notice). Dataset sim-v1-001, code 57cb0777ac.

ID	Hypothesis	Estimate	p	Holm p
H1	Higher AI accuracy increases reliance (0.90 vs 0.70 arm)	2.10	4.59e-05	0.000184
H2a	Displayed confidence increases acceptance (calibrated arm), holding AI correctness fixed	1.140	0.0449	0.0449
H2b	Displayed confidence increases acceptance (miscalibrated arm) — confidence heuristic	1.386	2.56e-08	1.28e-07
H4	Confidently-wrong AI produces more over-reliance than uncertain-wrong AI	1.777	0.000268	0.000805
H7	Higher over-reliance lowers complementarity in the 0.70 arm	-0.081	0.00263	0.00526
H6-exploratory	Appropriate reliance changes over trials (learning)	0.999	0.963	—

Table 3: Robustness of the acc90 (AI-accuracy) coefficient across specifications (SYNTHETIC).

Specification	OR	95% CI	p
Primary (disagreement trials)	2.10	[1.47, 3.01]	4.59e-05
Secondary (all trials + agreement dummy)	2.08	[0.93, 4.68]	0.0762
Independence working correlation	1.48	[0.99, 2.21]	0.0548
Drop first 5 trials/block	2.18	[1.56, 3.07]	6.43e-06
Drop RT $<$ 400ms	2.10	[1.47, 3.01]	4.59e-05

check that the corresponding test's false-positive rate is near its nominal level; scenarios contrasting two archetypes check that the estimator recovers the correct sign and rough magnitude of the difference. All validation is methodological: it establishes that the design and code can detect a known effect, and says nothing about whether that effect exists in people.

11 Discussion

No claim about human reliance behaviour is made in this draft. What is established is: (1) the experimental design is implementable end to end, from randomization through a browser-based trial sequence to a persisted, exportable dataset; (2) the outcome definitions and primary specification are precise enough to implement identically in two independent languages (TypeScript and Python), verified by a golden cross-language test; and (3) the analysis pipeline recovers known parameters from simulated data across eight qualitatively distinct participant archetypes, controls its false-positive rate under a null-confidence scenario, and achieves non-trivial power at the simulator's effect sizes with samples in the low hundreds. Once real pilot data exist, this section will be rewritten to interpret actual estimates against H1–H7, compare them to the related-work estimates in Section 2, and discuss where they agree or disagree with algorithm-aversion versus automation-bias accounts.

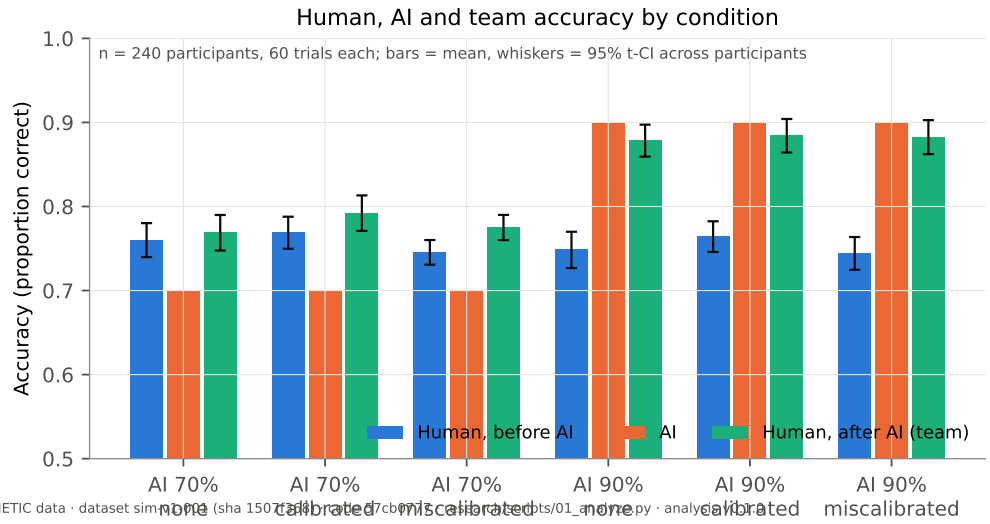

Figure 1: Initial human accuracy, AI accuracy, and final (team) accuracy by condition, computed from the archetype simulator (SYNTHETIC — see evidence notice above).

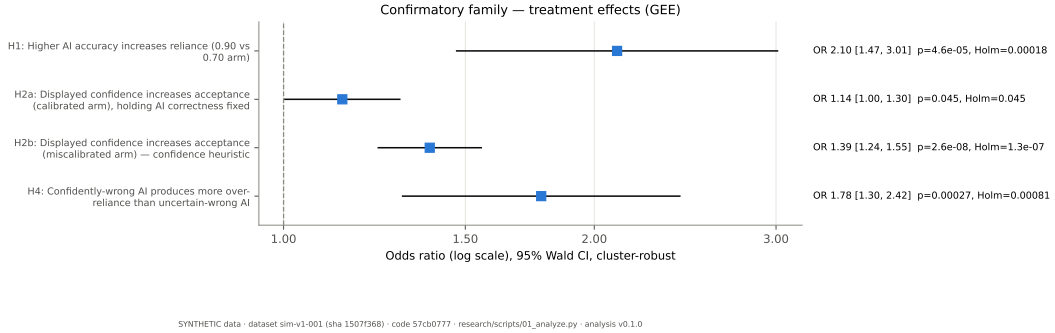

Figure 2: Confirmatory-family odds ratios with 95% cluster-robust confidence intervals and Holm-adjusted p-values, from the same simulated dataset. These are estimator-validation results, not evidence about human reliance.

12 Limitations

- **No human data.** This is the central limitation of the current draft; every quantitative result is a validation exercise on simulated participants, not a scientific finding.
- **Synthetic effect sizes are assumptions, not estimates.** Power and validation results use the simulator's built-in archetype parameters, which are engineering choices, not measurements; the pilot phase is required to re-estimate the between-participant variance before the confirmatory sample size is finalized (`docs/POWER.md`).
- **Two-task generalization.** Perceptual comparison and short-horizon forecasting are deliberately simple, expertise-free tasks; whether the same reliance patterns hold in higher-stakes or expertise-gated domains is not addressed here.
- **Within-session learning window.** 30 trials per task family may be too few to detect within-session learning even if it exists; experiment v2's feedback manipulation and any future

Table 4: Estimator validation scenarios against a participant simulator with known parameters.

Scenario	Expectation	Pass
S1_null_confidence	H2a and H2b p-values not small; OR close to 1	yes
S2_learners	early-window post-error adjustment negative (CI excludes 0)	yes
S3_nonlearners	post-error adjustment close to 0 (CI covers 0)	yes
S2_vs_S3	learners more negative than non-learners (early window)	yes
S4_vs_S5		yes
Type-I rate, null confidence (H2a/H2b, 16 seeds)	nominal 5%	12.5% / 0.0%
Rejection rate at N=120 (8 seeds)	H1/H2a/H2b/H4/H7	75%, 100%, 100%, 100%, 25%
Rejection rate at N=240 (8 seeds)	H1/H2a/H2b/H4/H7	100%, 100%, 100%, 100%, 75%

extension to more trials or a second session are natural next steps.

- **Wizard-of-Oz AI.** The AI is a controlled stochastic oracle, not a real model, which is standard for isolating accuracy and confidence effects but means results do not directly speak to any specific deployed system's calibration.

13 Ethics

No human-subject data have been collected under this protocol at the time of writing. The consent and debriefing text implemented in `web/src/app/experiment/consent` and `.../complete` discloses that the AI recommender is a controlled research instrument, not a deployed product, immediately in the debrief. Data collection is minimal by design: no names, emails, or IP addresses are stored; two demographic questions are optional; a participant code from a recruitment platform, if present, is stored only as a salted hash. Every participant receives a deletion key that permanently removes their participant, session and trial rows via a dedicated API endpoint, exercised by an automated end-to-end test. Before live collection begins this protocol requires IRB (or equivalent) review; nothing in this repository substitutes for that review.

14 Conclusion

We have specified and built, end to end, a reproducible platform for studying appropriate reliance on AI recommendations: a pre-registered two-experiment design, a browser-based repeated-decision engine shared byte-for-byte between the participant-facing application and an offline archetype simulator, a statistical analysis package whose every estimator is validated against simulations with known ground truth, and a manuscript pipeline that regenerates every number in this document from a hash-verified dataset. The only thing this platform does not yet contain is a human participant. That is the explicit next step.

References

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

Glenn W Brier. *Verification of forecasts expressed in terms of probability*. Monthly Weather Review, 1950.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 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*, 5(CSCW1):1–21, 2021.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1): 114–126, 2015.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70, 1979.

Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deceptive detection. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 29–38, 2019.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151: 90–103, 2019.

Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997.

Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 410–422, 2023.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

A Appendix

Codebook

See `docs/DATA_CODEBOOK.md` for the full 32-column trial record schema.

Reproducibility

`make reproduce` regenerates every artefact in this document from the canonical seed, recording the git SHA, dataset hash, and analysis version in `results/manifest.json`.

Literature matrix

See `docs/LITERATURE_MATRIX.md` for the full comparison table (citation, RQ, setting, sample, manipulation, outcome, result, limitation, relevance) underlying Section 2.