

Familiar Style and Calibrated Correction in Repeated AI Use

A framework for style recognition discrimination and substantive correction

Conceptual working paper

Version 0.1 | 17 October 2026

Luiza Scurtu

DOI: [10.5281/zenodo.23435113](https://doi.org/10.5281/zenodo.23435113)

Abstract

Repeated users of an artificial-intelligence agent may learn both its competence and how it sounds. This paper proposes that acquired style congruence can change acceptance independently of evidential quality, while any benefit to error management may arise through more than one route. Learned style–correctness contingencies can guide classification; reduced processing demands may also free attention for substantive checking when style itself is non-diagnostic. The framework distinguishes recognition, response criterion, classification sensitivity, competence learning, defect identification, successful correction, and final accuracy. Five competing mechanisms are specified: familiarity halo, anomaly vigilance, diagnostic cue learning, competence learning, and processing-efficiency learning. Building on appropriate-reliance research, calibration symmetry is treated as a requirement for identification and measurement: both familiar-looking wrong answers and unfamiliar-looking correct answers must be examined. It does not imply equal frequency, cost, or normative importance. The contribution is a bounded synthesis for acquired, source-specific style, not a new general principle of reliance or a demonstrated familiarity effect. Study 1 crosses acquired style congruence with correctness while preserving randomized assignment and measuring inferred source or version change. Study 2 varies cue diagnosticity and distinguishes classification through style from identifying and repairing substantive defects. Outcome feedback, trial-level learning, delayed unaided performance, and transfer are specified separately from assisted performance. Conversational affirmation and linguistic accommodation are adjacent pathways, not evidence of acquired-style effects. Governance implications remain provisional pending direct evidence.

Keywords: human-AI interaction; appropriate reliance; cue diagnosticity; familiarity; error detection; linguistic style; calibration

1. Introduction

Repeated interaction with an AI agent creates at least two kinds of learning. Users observe what the system gets right and wrong, forming beliefs about its competence. They may also learn recurring features of its expression: sentence rhythm, ordering conventions, degree of formality, characteristic transitions, or ways of presenting evidence. These signals can become recognisable even when the factual content changes. The central question of this paper is whether familiarity with that stable style improves or impairs people's ability to detect and correct AI errors.

Trust, advice use, error identification, successful correction, and final performance are non-equivalent outcomes (Lee & See, 2004; McBride et al., 2014). Appropriate-reliance research already distinguishes following correct and incorrect advice from achieving good final decisions (Guo et al., 2024; Schoeffer et al., 2025). An apparent benefit from familiar presentation might therefore reflect a shift in willingness to accept, useful cue learning, or better substantive checking. Calibrated correction requires those possibilities to be measured separately.

The review leads to a deliberately bounded claim. Research has established that people learn aspects of an aid's error boundary, that familiar or fluent material can shift truth judgements, that stylistic and social cues affect evaluations of artificial agents, and that inconsistency can lower reliance. Within this bounded search, no located experiment established the direction or existence of a causal style-familiarity effect on error discrimination after competence, task exposure, confidence, and clarity are separated.

Familiar style is modelled as a second-order cue to continuity with a learned source pattern. Its learned association with competence can guide reliance, while processing efficiency supplies a separate possible route to better checking. Both familiar-looking errors and unfamiliar-looking correct answers are needed to determine whether style changes discrimination or only willingness to accept. We call this calibration symmetry as a local research-design requirement, building on existing reliance theory rather than introducing a general theoretical discovery. Symmetry concerns identification and measurement; error frequencies, costs, and appropriate decision thresholds can be asymmetric.

Claim status. This is an integrative conceptual working paper with conditional, falsifiable predictions. It reports no original empirical results. The proposed studies are required to establish the focal effects, rather than a prerequisite for publishing the conceptual argument as such.

2. Constructs and evidential boundaries

2.1 Recognisable style and style familiarity

Recognisable style means a recurring pattern in the surface organisation and expression of answers that users can distinguish from alternative patterns above chance across new content. An assigned style can serve as a learned source signature without uniquely identifying an actual model. Features include syntax, discourse ordering, connective use, cadence, and stable rhetorical conventions. The definition excludes factual accuracy, task knowledge, latency, citation quality, and remembered performance. Confidence, clarity, warmth, and verbosity require separate controls because they can alter reliance independently.

Style familiarity is acquired recognition through exposure. It is not equivalent to prior use of ChatGPT, liking, source recognition, anthropomorphism, personal similarity, identity continuity, or knowledge of a system's

brand. These constructs may covary in ordinary use, but they cannot substitute for a manipulation check showing that participants learned the focal style.

This definition is intentionally narrower than “persona.” Classic computer-personality studies manipulated dominance, submissiveness, or introversion and measured attraction, satisfaction, or purchase intentions (Nass et al., 1995; Nass & Lee, 2001). Such evidence shows that people respond to stylistic consistency and similarity, but not that an acquired style changes truth discrimination. Contemporary model research also indicates that style may drift through interaction and vary substantially by genre (Frisch & Giulianelli, 2024; Rallapalli et al., 2026). A recognisable style is therefore an empirical property to be validated, not assumed from a prompt label.

2.2 Competence learning, task familiarity, confidence, and clarity

Learned competence is evidence accumulated about the agent's accuracy or conditional error pattern. Users can reasonably rely more on a system that has performed well, and feedback can teach where an aid is likely to fail (Bansal et al., 2019; Yin et al., 2019). Any naturalistic association between repeated use and reliance is therefore ambiguous unless performance history is held constant or manipulated independently.

Task familiarity is the user's practice with the domain, item structure, or decision rule. It may increase both unaided accuracy and the ability to evaluate advice. A style effect that appears only on rehearsed tasks may instead be task learning. Transfer to unpractised but structurally comparable items is required to distinguish the two.

Expressed confidence includes hedges, certainty markers, and assertive phrasing. Small changes in LLM uncertainty language can change agreement and confidence (Kim et al., 2024). *Clarity* includes readability, length, organisation, explanation detail, and apparent coherence. Explanations can increase reliance on both correct and incorrect LLM answers (Kim et al., 2025), while more interpretable presentations do not necessarily improve correction (Poursabzi-Sangdeh et al., 2021). Confidence and clarity must therefore be equated, independently rated, and modelled as manipulation-leakage checks rather than folded into style familiarity.

Outcome feedback supplies independently checked information about whether a decision was correct. It can teach task knowledge, the human's own competence, the AI's competence, or a conditional cue. An explanation describes or justifies an answer; without independent checking it is not outcome feedback. Immediate assisted accuracy, improved discrimination while a cue is present, and retained unaided learning are therefore distinct outcomes.

2.3 Detection, correction, and calibration

Error management comprises distinguishable stages: noticing that something may be wrong, locating the problem, generating or retrieving a correction, and adopting the improved final answer (McBride et al., 2014). A participant may flag a correct answer, correctly detect a wrong answer but fail to repair it, or accept a wrong answer after accurate initial work. The proposed outcome model separates:

- **classification sensitivity:** distinguishing wrong from correct AI answers, potentially using either substantive evidence or a learned stylistic cue;
- **response criterion:** willingness to flag or reject an answer, estimated separately from classification sensitivity;

- **successful substantive correction:** producing a demonstrably improved answer to wrong advice; preserving an already correct answer and adopting correct AI advice are recorded as distinct decision transitions;
- **final accuracy:** the task-level result after advice and any verification;
- **metacognitive calibration:** alignment between confidence and correctness; and
- **verification behaviour:** evidence inspection, latency, and optional checking.

Defect identification requires the participant to locate and explain the substantive error against a blinded scoring rubric. It is not established by classifying an answer as wrong, expressing distrust, or identifying the style. Trust, warmth, liking, and satisfaction remain secondary outcomes.

Discrimination alone does not define optimal reliance. A rational threshold also depends on available information, error prevalence, asymmetric losses, checking costs, and relative competence (Guo et al., 2024). Schoeffer et al. (2025) distinguish reliance from decision quality. Studies should specify payoffs and a benchmark using the information participants can access; correct and incorrect realised outcomes do not by themselves establish rational or irrational reliance.

3. Review method

3.1 Scope and search strategy

This is a structured narrative review designed to locate the closest causal evidence and test the novelty of the proposed configuration. It is not a systematic review and does not claim exhaustive retrieval. Searches were conducted across PubMed, publisher and DOI pages, ACM and ACL proceedings, arXiv, institutional repositories, and public scholarly web indexes. Search strands covered: (a) human–AI reliance, learned accuracy, automation errors, correction, and verification; (b) repetition, processing fluency, illusory truth, contradiction detection, belief bias, and signal detection; and (c) computer personality, conversational style matching, repeated interaction, persona consistency, anthropomorphism, and AI response congruence.

Inclusion was relevance-led: primary experiments, systematic reviews and meta-analyses, and computational work necessary to establish whether stable style is measurable. Preprints and dissertations were considered only where they addressed a close construct or current technical question and are labelled as such.

The audit did not independently reproduce every reported analysis or conduct an exhaustive correction-notice search. Access and verification limits are reported where consequential.

3.2 Selection logic and synthesis

Evidence was organised by the proposition it can support. Direct evidence for the focal claim would require acquired familiarity with a particular agent’s style, independent correctness variation, and behavioural identification or correction outcomes. No study meeting all three requirements was located within this narrative search. This bounded retrieval finding must not be read as “no prior study exists.” Adjacent findings are separated from proposed mechanisms and unknown effects.

4. What is established

4.1 Users can learn competence and conditional error patterns

Repeated experience with an automated adviser can teach users about its reliability. In a large randomised experiment, stated model accuracy and experienced performance both affected behavioural reliance, and experience partly displaced prior descriptions (Yin et al., 2019). Chen et al. (2018) similarly found that outcome feedback improved objective and subjective trust calibration in a phishing-detection task, whereas description alone primarily influenced subjective trust. These findings make competence history a mandatory control: repeated exposure does not produce “mere familiarity” in tasks where users also see outcomes.

Learning can be conditional rather than global. Bansal et al. (2019) showed that people formed more useful mental models when an AI's error boundary was parsimonious and less stochastic. Team performance depended on whether users could learn where the system failed, not simply on aggregate accuracy. Error type and distribution likewise produce state-specific patterns of under- and overreliance (Sanchez et al., 2014). A conspicuous error can also change later reliance differently depending on when it occurs and how much control users retain (Kim, Yang, & Zhang, 2023). These studies establish that experience can improve or distort reliance through learned performance structure. None isolates linguistic style.

The wider literature also rules out a simple expectation that experience always produces overtrust. Observing an algorithm err can cause algorithm aversion even when it remains superior to a human forecaster (Dietvorst et al., 2015), while other settings produce algorithm appreciation (Logg et al., 2019). Errors that violate a strong competence expectation may be especially damaging to trust (Madhavan et al., 2006). The direction of reliance depends on task, comparator, expertise, control, and error history.

Berger et al. (2026) reanalysed Vaccaro et al.'s (2024) 74 studies: ten supplied outcome feedback, none experimentally varied it, and only three examined change across trials. Feedback with explanations was associated with a positive synergy estimate, but uncertainty included zero. This motivates learning-sensitive designs, not a causal claim that feedback works or that learning is the most important variable.

Synergy is joint performance relative to the stronger partner alone. A gain in this measure need not demonstrate style recognition, substantive diagnosis, or retained unaided skill. The present studies accordingly specify feedback and learning measures while keeping their narrower acquired-style estimands.

4.2 Familiarity and fluency often raise acceptance, but cue use is learned

Repetition reliably increases perceived truth. A 2026 three-level meta-analysis synthesised 182 studies, 366 effects, and 31,184 participants; after adjustment for small-study effects, the average illusory-truth effect was modest ($g = 0.37$, 95% CI [0.30, 0.44]) and heterogeneous (Ye et al., 2026). An earlier meta-analysis of 51 studies reached the same broad conclusion (Dechêne et al., 2010). Repetition can raise perceived truth even when people possess knowledge that contradicts the claim (Fazio et al., 2015), and prior exposure can increase perceived accuracy of false news headlines (Pennycook et al., 2018).

This evidence is a serious warning against assuming that familiarity improves checking. It is nonetheless indirect. These paradigms typically repeat propositions, manipulate perceptual fluency, or measure truth ratings. They do not train recognition of an agent's generative style and then observe active diagnosis and correction of new factual content.

Fluency is not inherently a truth signal. Unkelbach (2007) reversed the usual relationship by teaching an environment in which fluent items were more likely to be false. This supports learned cue interpretation: a recognisable style might be used when it predicts error. Diagnostic cue use is one possible source of value. A separate possibility is that a familiar presentation reduces processing demands and leaves more capacity for checking the content, even without a style–correctness contingency. That processing-efficiency hypothesis is not established by the cited fluency experiments.

Familiarity can be separable from source recollection. Repeated statements attributed to non-credible sources were later judged truer unless the source was especially memorable (Begg et al., 1992). Ease of retrieval can also inflate confidence in general-knowledge answers, including related incorrect answers (Kelley & Lindsay, 1993). The implication is not that a recognisable AI style must create an illusion of truth. It is that confidence and source-memory effects can mimic better knowledge and require independent measurement.

4.3 Surface congruence can hide errors and anomaly can create false alarms

Contradiction-detection research demonstrates that people sometimes miss distortions that partially match stored knowledge. In the Moses illusion, a question such as “How many animals of each kind did Moses take on the ark?” often elicits “two,” despite the name mismatch. Signal-detection analyses indicate that some manipulations alter response criterion without improving sensitivity to the distortion (Kamas et al., 1996). Expertise reduces but does not eliminate these failures (Cantor & Marsh, 2017), and undetected contradictions can later contaminate memory (Bottoms et al., 2010). Text predictability and lexical familiarity can also impair proofreading under some conditions (Pilotti et al., 2009).

These findings provide an analogue for familiar-looking wrong answers: an expected surface pattern can support a gist-consistent response while a local error escapes notice. The analogy must remain explicit. A misleading question is not a familiar AI agent, and semantic overlap is not identical to style recognition.

The converse identification problem also matters. Believable conclusions are more readily endorsed in reasoning tasks, yet a hierarchical Bayesian signal-detection synthesis found that believability generally shifted endorsement criterion rather than underlying discrimination (Trippas et al., 2018). Rejecting an unusual-but-correct answer may therefore reflect a decision threshold rather than diminished sensitivity. Raw accuracy or challenge rate cannot identify the difference; the costs of false rejection and false acceptance need not be equal.

Attempts to improve reasoning through superficial disfluency are not a secure foundation. Song and Schwarz (2008) reported that a difficult font reduced Moses errors but also impaired answers to undistorted questions. Later tests failed to reproduce a reliable benefit (Janoušková et al., 2022). An accuracy focus or explicit expectation that falsehoods may occur has more consistent support as a boundary condition (Brashier et al., 2020; Jalbert et al., 2020). These interventions affect goals and vigilance, not learned style.

4.4 Presentation affects reliance, but trust is not calibration

AI explanations and conversational formats change how advice is received. In a preregistered LLM study with 308 participants, explanations increased reliance on both correct and incorrect responses; clickable sources reduced reliance on incorrect responses (Kim et al., 2025). The paper also found lower reliance when explanations contained natural inconsistencies, but the inconsistency analysis capitalised on naturally occurring features rather than a fully orthogonal style manipulation. It therefore suggests an anomaly mechanism without establishing that anomaly reliably improves discrimination.

A newer experiment crossed LLM-answer correctness with consistency in a factual question-answering task. Inconsistency reduced reliance even when the final answer was correct and did not improve participant accuracy overall; the effect also varied with how the LLM was framed (A. Chen et al., 2026). This is a boundary case rather than direct style evidence. A correct answer accompanied by mutually incompatible supporting statements is substantively defective, not merely sound advice expressed in an unfamiliar voice. The study supports the narrower proposition that conspicuous anomalies can suppress acceptance despite a correct final answer. It does not establish false rejection caused by stylistic unfamiliarity.

Bao et al. (2025) directly connected textual style explanations to reliance in AI-assisted authorship attribution. Example-based style rewrites supported more appropriate reliance, whereas presenting both style rewrites and feature rationales increased overreliance; no explanation condition improved overall accuracy. The study concerns style evidence used to explain authorship, not familiarity acquired with a recurring source style. It therefore narrows the adjacent literature without establishing the focal effect.

Other work reinforces the need to separate apparent comprehensibility from successful correction. Transparent models can be easier to simulate without enabling users to correct large errors (Poursabzi-Sangdeh et al., 2021). Cognitive-forcing interfaces reduce following of wrong AI advice but can impose usability costs and do not automatically improve team performance (Bućinca et al., 2021). Across five preregistered personnel-selection experiments, explanations did not reliably repair the performance harm caused by incorrect AI advice (Cecil et al., 2024). Conversely, coherent explanations for why an aid might err can increase trust and reliance even when greater reliance is unwarranted (Dzindolet et al., 2003).

4.5 Agent style is perceivable, relational, and context-sensitive

Humans respond socially to artificial communication styles. Users recognised dominant or submissive computer personalities and preferred matched personalities (Nass et al., 1995); synthetic voice personality and voice–text consistency affected evaluations and purchase behaviour (Nass & Lee, 2001). Conversational style-matching and linguistic-alignment studies examine user evaluations and workload (Hoegen et al., 2019; Spillner & Wenig, 2021). These findings concern social or processing outcomes, not acquired source-specific style and factual error correction.

Value-congruent framing can increase endorsement and willingness to pay (Hirst et al., 2026). Feeling understood is a reported pathway to commercial engagement; endorsement is linked to making a more compelling case. This is neither acquired style familiarity nor demonstrated calibration. Eder et al. (2026) found more favourable trust and persuasiveness ratings for opinion-aligned agents, with weak or null personality-alignment effects in several comparisons; those ratings do not establish successful behavioural persuasion. Human-like communication can soften negative user-experience responses to chatbot mistakes (de Sá Siqueira et al., 2024). None of these studies isolates the focal style mechanism.

Social categorization may also shape evaluations when an AI provides no explicit visual, vocal, or identity cues ordinarily used to infer social identity. Across five principal and eight supplementary studies, Gamez-Djokic and Waytz (2026) found that participants frequently associated AI with Whiteness. Perceiving an AI as more White was associated with greater trust and perceived persuasiveness, while an experiment manipulating the stated racial composition of training data affected the cognitive demands of tasks assigned to otherwise identical systems. The trust and persuasiveness findings are associative; the experimental result concerns task allocation, not reliance on correct or incorrect advice. This research therefore identifies an adjacent social-evaluative pathway through which perceived competence or authority may vary without demonstrated

capability differences. It does not test acquired source-specific style, repeated exposure, correctness discrimination, or correction, and should not be treated as evidence for the focal familiarity effect.

Nor is agent style necessarily stable across contexts. Computational work finds measurable differences across models, while genre can explain more stylistic variation than source and decoding settings also matter (Rallapalli et al., 2026). Persona-prompted LLM agents can align linguistically during interaction, reducing the recognisability of their initial profiles (Frisch & Giulianelli, 2024). Consequently, an apparent style anomaly in deployment may signal a change of task, prompt, retrieval context, policy, or model rather than a factual error.

Accommodation introduces an additional distinction. Blevins et al. (2026) found linguistic convergence in model completions of human dialogues; this is evidence that model output adapts to context, not an experiment on user familiarity or correction. P. Chen et al. (2026) report progressive bidirectional convergence in observational human–AI conversations. In a separate large-scale observational preprint, Blevins (2026) found that models overconverged toward users, whereas human-to-model convergence was broadly comparable with human–human baselines. The findings differ in emphasis rather than directly contradicting one another, and neither identifies an effect on factual checking. AI adaptation toward a user, user adaptation toward an AI, and acquired recognition of a stable source style remain distinct processes. The fixed A/B styles below isolate the last process under controlled conditions.

Osler (2026) proposes that apparently responsive dialogue can support validation and deference, including outside clinical settings. Such an account makes a relational pathway plausible even where a user does not literally believe the AI has a mind. It does not demonstrate a causal familiarity effect. Model agreement at the expense of truth is also documented in sycophancy evaluations (Sharma et al., 2024). These findings require separation of validation and endorsement from acquired style, rather than treating all agreeable or personally resonant output as familiar.

5. The diagnosticity–calibration framework

Recognisable style may retrieve expectations about a source, task region, or production process. It may also reduce the effort needed to parse a response, affecting the capacity available for checking. These are different pathways: style can serve as evidence about an answer or change how efficiently that answer is inspected. The framework separates three stages without assuming that they identify a single mechanism:

1. **Recognition:** the current response is experienced as congruent with, or deviant from, a learned stylistic signature.
2. **Interpretation and processing:** congruence or deviation retrieves learned expectations, changes processing demands, or suggests a source or system change.
3. **Epistemic action:** the user accepts, verifies, challenges, or corrects the answer under constraints of knowledge, effort, incentives, and available evidence.

Familiarity need not have a single direction. A familiar surface may generalise trust, retrieve a useful error cue, or reduce presentation-processing demands. A deviation may prompt vigilance or a rational inference that the source changed. Five competing mechanisms are summarised in Table 1; they can coexist and are not an exhaustive taxonomy.

Explanation	Distinguishing observable pattern
Familiarity halo	Congruent style raises acceptance of correct and wrong answers; response criterion changes without better discrimination.
Anomaly vigilance	Deviant style increases checking or rejection of wrong and correct answers; hits on wrong advice and false alarms on correct advice rise together.
Diagnostic cue learning	Classification follows the learned style–correctness contingency; substantive defect identification need not improve.
Competence learning	Reliance follows experienced performance history or task region rather than surface style.
Processing-efficiency learning	With non-diagnostic style, substantive checking or repair improves with reduced processing burden; latency alone is insufficient.

Table 1. Competing mechanisms and their predicted observable patterns. More than one mechanism may operate in the same interaction.

Mean trust or aggregate accuracy cannot identify these mechanisms. The relevant questions are what the user learned, whether a cue remains valid, whether substantive checking improves, and whether effects follow style or remembered competence. The comparisons below specify distinguishing patterns and limits rather than unique signatures that prove a mechanism.

5.1 Learned diagnosticity

Established premise: people can learn conditional error patterns and interpretations of fluent cues (Bansal et al., 2019; Unkelbach, 2007).

Hypothesis: a stable style can become a diagnostic cue when its relation to correctness is learned. This can improve classification without improving substantive error identification.

Required evidence: behaviour follows the randomized cue-correctness contingency, including reversal or extinction when its validity changes. Claims of better error diagnosis additionally require defect identification and repair on content-based probes.

5.2 Fluency or identity halo

Established premise: repetition and fluent presentation can increase truth judgements or reliance (Fazio et al., 2015; Kim et al., 2025; Ye et al., 2026).

Hypothesis: familiar style may lower the threshold for accepting both correct and wrong answers, without improving sensitivity.

Required evidence: an acceptance shift across correctness levels with criterion and sensitivity estimated separately. Its absence in a sufficiently precise randomized comparison weakens this proposed halo pathway.

Relational continuity, feeling validated, or attributed epistemic authority may contribute to this halo even when warmth and confidence are matched. They are candidate pathways within the present account, rather than a sixth identified mechanism. Measure them after primary outcomes, and use a separately randomized intervention if a causal mediation claim is intended. Their association with acceptance cannot establish the pathway.

5.3 Expectancy violation and vigilance

Established premise: inconsistencies and expectation-violating errors can reduce reliance (A. Chen et al., 2026; Madhavan et al., 2006).

Hypothesis: unusual style may prompt checking or rejection of both wrong and correct answers.

Required evidence: increased checking or rejection in the alternative style, including false alarms on correct answers. Increased substantive sensitivity instead would require a checking-efficiency or diagnosticity account. Error prevalence, costs, and user knowledge constrain the predicted net benefit.

5.4 Competence and task learning

Established premise: observed accuracy and feedback change reliance and learned error boundaries (Bansal et al., 2019; Yin et al., 2019).

Alternative explanation: effects attributed to style reflect remembered competence or task practice.

Discriminating evidence: under an independent history manipulation, reliance follows the learned accuracy history after restyling; with histories equated, the apparent style effect diminishes. Transfer items and unaided baselines distinguish task learning. Study 1 controls this rival but cannot fully estimate it without an additional history contrast.

5.5 Processing efficiency learning

Established premise: linguistic alignment studies report user-experience and workload outcomes (Spillner & Wenig, 2021), but they do not establish that acquired source-specific style improves error diagnosis.

Competing hypothesis: practice parsing a stable style reduces presentation-processing demands, leaving capacity for substantive checking when time or attention is constrained. Style need not predict correctness.

Discriminating evidence: improved defect identification or repair under non-diagnostic style, alongside lower processing burden and a favourable speed–accuracy trade-off. An independently randomized load or time-budget contrast tests the proposed resource pathway. Faster acceptance without better substantive performance supports efficiency of responding, not better oversight.

5.6 Confidence clarity and social response

Established premise: uncertainty expression, explanations, warmth, and conversational framing influence reliance and evaluations (Anderl et al., 2024; Kim et al., 2024, 2025).

Alternative explanation: the “familiar” style is simply clearer, more assertive, warmer, shorter, or more genre-conventional.

Discriminating evidence: content-preserving restyling, independent ratings, and exclusion or redesign of stimuli that differ materially on these dimensions.

5.7 Personalization and style familiarity

A response conditioned on a user profile may be personally relevant, persuasive, or experienced as understanding without being stylistically familiar. Tailored content does not establish acquired recognition of a stable source pattern. This paper therefore reserves **style familiarity** for demonstrated recognition learned through exposure. It reserves **representation-mediated familiarity** for the narrower longitudinal case in which repeated profile-conditioned interaction produces a recognisable recurring pattern that the user has demonstrably acquired.

A persistent representation introduces another problem: subjective fit may increase without greater fidelity to the person. The companion working paper, *Reciprocal Miscalibration in Personalized AI Correction*, distinguishes static dual miscalibration from a hypothesised feedback process over time. Here the distinction prevents personalized relevance or feeling understood from substituting for acquired style recognition.

6. What the framework contributes

The literature supports a bounded integration for acquired, source-specific style. It does not establish that familiar AI style improves correction or causes overreliance. Three established distinctions guide the contribution: correctness versus reliance, classification versus substantive diagnosis, and discrimination versus expected decision value.

First, source-specific style recognition must be distinguished from competence history, task familiarity, perceived source identity, and general readability. The contribution concerns their separation in repeated AI use, rather than asserting that prior reliance research ignored experience or calibration.

Second, calibration symmetry specifies a design and reporting requirement for this configuration. Both familiar-wrong and alternative-correct trials must be included, alongside the other two cells. This applies established distinctions among overreliance, underreliance, and decision quality (Guo et al., 2024; Schoeffer et al., 2025). Raees and Papangelis (2026) review continuing differences among appropriate-reliance constructs. Symmetry neither predicts equal error rates nor assigns equal consequences to the two mistakes.

Third, the framework compares five routes to similar headline outcomes. A halo shifts acceptance; vigilance shifts checking; diagnostic learning supports cue-based classification; competence learning tracks performance history; and processing-efficiency learning may improve substantive inspection even when style predicts nothing about correctness. Measuring classification alone cannot distinguish the last three routes.

The governance implication is provisional. Retaining a familiar style after competence changes may retrieve obsolete expectations, while changing style alone may imply a new source. Explicit version and provenance signals are candidate design responses to be evaluated, not remedies established by the framework.

The synthesis applies signal detection and appropriate-reliance distinctions to an acquired-style $\times$ correctness design. It adds explicit checks for inferred source change and separates cue-based classification from substantive defect identification and repair. This is the proposed contribution; the correctness matrix and criterion–sensitivity distinction are not claimed as new theoretical principles.

The contribution is a bounded synthesis and test design for acquired source-specific style. Existing accounts already address conversational validation and human learning; linguistic-accommodation research also limits assumptions of style stability. The specific causal effect proposed here remains open. Established component findings motivate the propositions; they do not confirm them.

Five falsifiable propositions correspond to the competing mechanisms. Their boundary conditions and disconfirming observations are part of the claims:

P1 Familiarity halo. With non-diagnostic style and matched performance histories, congruence increases acceptance of correct and wrong answers without a sensitivity gain. A precise absence of an acceptance shift weakens this route; improved defect identification requires another mechanism.

P2 Anomaly vigilance. Alternative style increases checking or rejection, including false alarms on correct advice, particularly when source continuity is uncertain. No checking or rejection increase under a recognised deviation weakens this prediction. Whether the response improves expected value depends on costs and error prevalence.

P3 Diagnostic cue learning. With marginal accuracy and style frequency held equal, classification follows a learned style–correctness contingency and adjusts when that contingency reverses or disappears. This effect need not transfer to defect identification on style-masked probes. No contingency-sensitive learning weakens this route.

P4 Competence learning. When performance history is independently varied and style is swapped, reliance follows the accuracy history more than the current surface form. Persistence of an acquired-style effect with histories equated limits this explanation; the added history contrast is required to test its positive prediction.

P5 Processing-efficiency learning. With non-diagnostic style, matched competence history, and sufficient task knowledge, acquired congruence improves substantive defect identification or repair by reducing presentation-processing demands. The benefit should be detectable under a preregistered level of moderate time or attention constraint and diminish when ample time removes that bottleneck. Equal performance with only faster acceptance, or gains solely on cue-present classification, weakens the proposed checking benefit. Extreme load and absent knowledge are boundary conditions rather than predicted benefits.

7. Future falsification studies

Study 1 estimates acquired style congruence with cue diagnosticity absent and competence history matched. Study 2 estimates cue use under a randomized style–correctness contingency. Neither experiment alone uniquely identifies all five mechanisms. Additional load or competence-history contrasts should be separately powered and preregistered when the corresponding mechanism is the target.

7.1 Common stimulus and sampling requirements

Develop two distinctive but non-caricatured styles, A and B, using discourse ordering, connective preferences, and sentence cadence. They must not differ systematically in confidence, hedging, warmth, politeness, verbosity, citation count, explanation depth, visual formatting, or genre conventionality. Generate one canonical answer for each item, create content-preserving restylings, and verify semantic equivalence manually and with independent ratings.

An independent sample should establish recognition above chance and practical equivalence in global clarity, confidence, readability, plausibility, warmth, and explanation quality using prespecified equivalence bounds. Non-significant differences alone do not establish equivalence. Acquisition-related processing efficiency remains an outcome, not a nuisance to regress away. Verify correctness independently, blind semantic checks to condition, and norm content conventionality separately from surface style.

Recruit participants with measurable, non-ceiling competence in an objective domain and sample many independent items. Set sample size by simulation for a preregistered smallest effect of interest, allowing participant and item variance, planned interactions, and missingness estimated from pilot data. No final recruitment target is justified by the present conceptual paper.

7.2 Study 1: Acquired style congruence

Study 1 estimates the causal effect of congruence with an acquired stable style on classification, criterion, substantive correction, and final accuracy. Stable exposure also makes that style expected, so this is not an estimand of familiarity purified of expectancy. A positive result can motivate a mechanism study without establishing one.

At baseline, participants answer separate task items unaided and report confidence. During acquisition, each interacts with an agent consistently using A or B, counterbalanced between participants. Exposure count, task mix, content, timing, source label, feedback, and objective performance history are identical between style assignments. Correctness is independent of style during acquisition and test. There is no varied-style group or diagnosticity manipulation in this study.

At test, hold the source label constant and cross answer correctness (correct or wrong) with style congruence (acquired or alternative). Counterbalance items so that each proposition appears in every cell across participants. Include trained-task and transfer-task families, but do not make style predictive of correctness. The four diagnostic cells are congruent-correct, congruent-wrong, alternative-correct, and alternative-wrong. The two decisive cases remain:

- **familiar-looking wrong answers**, where acquired congruence may suppress checking; and
- **unusual-but-correct answers**, where the alternative style may provoke unwarranted rejection.

Holding the displayed source label constant does not guarantee perceived identity continuity. Measure whether participants infer a different source, model, system version, or competence level, preferably after the primary outcome block or in a separate validation sample. Such inferences may rationally change reliance. Report them as potential pathways and test a source-continuity disclosure in a separate randomized contrast if needed. Do not casually adjust them away as covariates: they can be consequences of the style manipulation.

During acquisition, use identical independently checked correctness feedback after the participant's final decision; fix its timing and explanatory detail across A/B assignments. State whether feedback scores the initial answer, the AI answer, and the final answer. During the primary test block, withhold correctness feedback until the block ends and score judgments before any disclosure. This defines a test of acquired congruence under a specified learning history, not a general estimate across all feedback regimes. Provide verification access equally across cells.

After the primary block, measure perceived validation, interpersonal continuity, and epistemic authority separately from recognition and inferred source or version change. Keep these post-treatment measures descriptive. Independent pretests should also examine whether restyling changes apparent endorsement of the user's views, which can alter acceptance without acquired familiarity.

7.3 Outcomes, analysis, and falsification

On each trial record an unaided answer, confidence, AI response, correctness judgement, optional evidence inspection, challenge, proposed correction, final answer, final confidence, and time. Preregister classification sensitivity and final-answer accuracy as primary endpoints, with criterion reported alongside sensitivity. Score defect identification and successful substantive repair independently using blinded rubrics and inter-rater agreement. Distinguish preserving correct initial work, correcting initially wrong work, and harmful switching.

Secondary outcomes include verification effort, workload, confidence calibration, remembered reliability, recognition, and trust.

Use a preregistered hierarchical signal-detection or equivalent response model with participant and item effects and prespecified congruence $\times$ correctness contrasts. Define “wrong answer” as the signal: flagging wrong advice is a hit; flagging correct advice is a false alarm. Report hit and false-alarm rates by style, as well as sensitivity and criterion under stated model assumptions. Analyse all randomized participants in their assigned acquisition condition, including those who fail recognition checks. Recognition-defined and per-protocol results are secondary; complier causal estimates require an appropriate design and explicit identification assumptions. Post-treatment confidence, clarity, recognition, and identity judgements are not routine adjustment covariates.

An epistemic benefit requires better discrimination, substantive repair, or decision value, rather than greater trust alone. Increased acceptance across correctness levels supports a halo or criterion account; increased rejection and false alarms support vigilance. With style non-diagnostic, a sensitivity or repair gain is consistent with processing efficiency but does not establish that mechanism. Test P5 with an independently randomized load contrast and a joint speed–accuracy analysis, not latency alone. A precise null weakens only the prespecified effect within the sampled task, styles, and exposure dose.

Study 1 cannot positively identify competence learning because it holds history constant. A separately powered competence-history $\times$ current-style contrast tests P4. Likewise, source identity, fluency, and workload associations are descriptive unless their causal pathways are independently manipulated. The primary intention-to-treat effect remains valid as an assignment effect even when the recognition manipulation is weak; the interpretation then concerns that delivered exposure, not successfully acquired familiarity.

Preregister trial or block trends as secondary learning outcomes, including planned early–late contrasts rather than selecting the best-performing period after inspection. Use separate, counterbalanced item sets for a delayed unaided assessment and transfer to new content with the stylistic cue masked. Do not provide answer feedback during those assessments. Record delayed-test attendance and missingness. This separates immediate assisted performance, style-dependent classification, and retained content knowledge; within-person gains alone do not identify a feedback effect.

7.4 Study 2: Cue-diagnostics learning and transfer

Study 2 tests learning to use style as an error cue. Retain a common acquisition stage in a stable assigned style, then give all groups the same proportions of acquired and alternative styles with identical, independently checked correctness feedback after each completed decision. Hold explanation availability and detail constant across contingency arms. Random assignment varies the contingency only: diagnostic, non-diagnostic, or reverse-diagnostic. Marginal correctness and style frequency must match. For example, with both margins at 50%, the probabilities of an error given a deviation can be .75, .50, and .25, with the complementary probabilities for congruent responses. Pilot the exposure needed for learning and preregister it.

An initial test preserves the contingency; a separately labelled neutral-transfer phase removes it to measure persistence and extinction. Report classification and reliance discrimination, successful substantive correction, final-answer accuracy, and defect identification separately. A participant can classify by style without

inspecting the answer. Include new, content-equated probes with stylistic cues removed or normalized, scored for locating the actual defect and producing its repair. Report cue-present and cue-masked performance separately, including within-style correctness discrimination. Cue-based gains without gains on those probes support diagnostic cue use, not enhanced capacity to detect reasoning errors.

Preregister learning curves and adaptation after the contingency changes. Give no correctness feedback during the initial cue-masked diagnostic probes, so their score reflects knowledge available before the answer is disclosed. Add a delayed unaided test on new items to assess retention. Success during the feedback-rich acquisition phase alone is not evidence that participants can later inspect and repair substantive errors without the cue.

7.5 Pre-registration and openness

Before collection, define estimands, primary and secondary endpoints, smallest effects of interest, stopping and attrition rules, exclusions, error prevalence, payoff structure, item sampling, multiplicity control, and confirmatory models. Preserve assignment in primary analyses. Score reliance principally on initial human–AI disagreement trials; when answers already agree, final agreement alone cannot establish reliance. Archive stimuli, correctness standards, transformations, and consent-compatible data and code. Obtain required ethics approval and consent before recruitment. Keep the two studies distinct.

7.6 A separate learning opportunity module

A separately powered experiment can randomize outcome-feedback availability and explanation availability while holding style and objective accuracy fixed. The feedback must be independently validated; explanations must not reveal an answer key in only one arm. Prespecify gains in unaided knowledge, discrimination, and expected decision value across trials and at a delayed transfer test. This tests learning opportunities, not the acquired-style hypothesis. A claim that feedback changes the style effect requires an additional randomized style contrast and power for that interaction. Neither Study 1's matched feedback history nor Study 2's cue contingency identifies the causal effect of providing feedback.

8. Material weaknesses and boundary conditions

Several issues materially weaken the case for a strong conceptual or practical claim.

1. **Direct evidence remains limited.** The narrative search located no experiment meeting the focal configuration; the paper reports no observed causal style-familiarity effect and makes no exhaustive absence claim.
2. **Style is a fragile manipulation.** Even careful restyling may alter semantic emphasis, apparent confidence, warmth, conventionality, or evidence quality. A failed equivalence check invalidates a pure familiarity interpretation.
3. **Familiarity has inseparable rivals.** Stable exposure creates expectancy as well as familiarity; alternative style can also imply a new model or competence regime. Measure those inferences, preserve assignment, and limit interpretation to acquired style congruence unless an additional contrast identifies the mechanism.
4. **Styles drift and accommodate.** Genre, prompts, retrieval, updates, and both interlocutors' linguistic adaptation can change surface form. The A/B design estimates acquired congruence with controlled styles; naturalistic persistence of that signature requires separate validation.

5. **Ground-truth tasks are selective.** Objective questions permit validation but do not represent open-ended professional judgement. Generalisation to medicine, law, management, or creative work must be earned separately.
6. **Human knowledge sets the ceiling.** Participants cannot reliably diagnose subtle errors without relevant knowledge or usable external evidence. Low knowledge may turn the task into trust measurement.
7. **Anomaly and difficulty can be confused.** Counterintuitive correct answers must be normed independently of stylistic deviation, or false rejection may reflect substantive implausibility.
8. **Monitoring has costs.** Error prevalence, asymmetric losses, verification cost, and relative competence affect the appropriate threshold. Discrimination and final accuracy are informative components rather than a complete definition of optimal reliance.
9. **Demand characteristics are plausible.** Recognition checks and obvious style swaps may reveal the hypothesis. Covert counterbalancing and delayed checks are preferable.
10. **The current evidence base is mostly brief and artificial.** Durable effects in long-running professional relationships with evolving agents remain unknown.
11. **Learning depends on evidence.** Feedback availability, timing, reliability, task knowledge, and opportunities to assess relative competence constrain what can be learned. Findings from no-feedback tasks cannot establish an enduring incapacity for calibrated reliance.
12. **Relational validation is an adjacent pathway.** Feeling understood or affirmed may influence acceptance without acquired recognition, and clinical feedback accounts do not establish the focal effect or its prevalence.

These boundaries specify what evidence must show before a result about acquired style can guide design. Cue-based classification and substantive error diagnosis must not be collapsed even when both improve aggregate performance.

9. Implications for design and governance

No current evidence supports prescribing a stable recognisable style as a way to improve human oversight. Nor does it support suppressing style consistency on the assumption that familiarity necessarily creates complacency. The framework instead treats style as a possible metacognitive cue whose validity may change across tasks and model updates.

The framework motivates provisional practices for testing. Evaluation should include discrimination, false alarms, repair, and final outcomes alongside trust. Direct version and provenance signals may help when an old style masks a competence change or a new style suggests one, but their effects on reliance, workload, and decision value remain to be established. These are testable design implications rather than empirically validated prescriptions.

The present synthesis already supports one evaluation principle: do not judge style interventions from trust, challenge frequency, or performance on wrong-answer trials alone. The proposed study could inform a later design choice only if effects survive transfer tasks and nuisance controls. Even then, the goal would be calibrated use, not maximum familiarity or maximum scepticism.

10. Conclusion

Acquired style congruence may shift acceptance, trigger vigilance, support diagnostic cue use, or reduce processing demands for checking. Competence and task learning remain alternative explanations. None of these possibilities is established for the focal repeated-interaction configuration by the present synthesis.

The contribution applies established reliance and signal-detection distinctions to source-specific style while separating criterion, classification sensitivity, competence learning, and substantive correction. Calibration symmetry is an identification and measurement requirement, not a claim that errors have equal prevalence or cost. Style's predictive validity is one route to epistemic value; processing efficiency is a separate conditional hypothesis.

Future tests should preserve randomization, document independent outcome feedback, measure inferred source change and relational interpretations, and report defect identification and repair alongside classification and final accuracy. Trial-level change, delayed unaided assessment, and cue-masked transfer distinguish learning from assisted performance. Neither philosophical conversational-feedback accounts nor the learning reanalysis determines the direction of an acquired-style effect. Governance implications remain provisional.

Acknowledgements and provenance

This conceptual working paper was developed within Life Laboratory's research process. AI assistants were used for literature synthesis, drafting and revision. Luiza Scurtu retains responsibility for final interpretation and publication. The paper reports no original participant study or empirical dataset.

Appendix A. Concise evidence register

Source	Design or evidence type	What it supports	Why it does not answer the focal question
Bansal et al. (2019)	Repeated human–AI task with learnable versus stochastic error boundaries	Users can learn where an AI fails; error-boundary structure matters	No style manipulation; learning concerns competence
Yin et al. (2019)	Large randomised studies of stated and experienced accuracy	Accuracy history changes behavioural reliance	Repetition and competence are intentionally coupled
Ye et al. (2026)	Meta-analysis, 182 studies, 366 effects, $N = 31,184$	Repetition increases perceived truth on average, with heterogeneity	Repeated content and truth ratings, not agent style or correction
Unkelbach (2007)	Two fluency-learning experiments	Learned cue validity can reverse the truth effect	Perceptual fluency, not AI style
Kamas et al. (1996)	Moses-illusion experiments with signal detection	Surface match can produce criterion effects without sensitivity gains	Misleading questions, not repeated AI interaction

Source	Design or evidence type	What it supports	Why it does not answer the focal question
Trippas et al. (2018)	Hierarchical Bayesian ROC meta-analysis	Believability mainly shifts endorsement criterion	Syllogistic reasoning, not agent familiarity
Kim et al. (2025)	Think-aloud study and preregistered experiment, N = 308	Explanations raise reliance on correct and wrong LLM answers; sources and inconsistent explanations can reduce reliance on errors	Naturally occurring and substantive inconsistencies, not acquired stylistic unfamiliarity
A. Chen et al. (2026)	Two experiments; correctness and consistency examined in factual QA	Inconsistency can lower reliance even when the final answer is correct	Contradictory support is a substantive defect; boundary case rather than style evidence
Poursabzi-Sangdeh et al. (2021)	Preregistered interpretability experiments, total N = 3,800	Apparent interpretability need not improve error correction	Model transparency, not style familiarity
Buçinca et al. (2021)	Cognitive-forcing experiment	Reduced overreliance can carry usability costs and need not improve team performance	One short task; no style familiarity
Nass et al. (1995); Nass and Lee (2001)	Computer personality and consistency experiments	People perceive and evaluate artificial styles socially	Attraction/evaluation outcomes; no answer correctness
Hirst et al. (2026)	Preregistered value-congruent chatbot experiments	Congruence can increase endorsement and willingness to pay; feeling understood is a reported pathway to commercial engagement	Value alignment, not acquired style; persuasion, not calibration
Frisch and Giulianelli (2024)	Agent-only persona and alignment study	Interaction can erode persona recognisability	No human users or correctness outcome
Rallapalli et al. (2026)	Computational preprint, 467,985 texts	Style varies measurably by model, genre, and decoding	No human familiarity, trust, or correction

Source	Design or evidence type	What it supports	Why it does not answer the focal question
Anderl et al. (2024)	Conversational presentation crossed with information accuracy	Presentation mode can reduce accuracy discrimination	Conversational format bundled; no acquired style familiarity
de Sá Siqueira et al. (2024)	Human-like style crossed with chatbot mistakes	Social style changes user experience around mistakes	Style bundle; UX rather than correction accuracy
Guo et al. (2024); Schoeffer et al. (2025)	Formal reliance and decision-quality frameworks	Following advice, signal discrimination, and final value are separable	No acquired source-style manipulation
Raees and Papangelis (2026)	Narrative review of measurement constructs [Preprint]	Trust, reliance, and appropriate-reliance metrics require distinction	No causal test of acquired style or processing efficiency
Berger et al. (2026); Vaccaro et al. (2024)	Reanalysis of collaboration studies; original meta-analysis	Learning opportunities and change across trials deserve explicit evaluation	Randomized feedback effects, retained skill, or an acquired-style effect
Osler (2026)	Philosophical conversational-feedback account	Validation and deference are close alternative interpretations	A causal style-familiarity effect or clinical incidence
Sharma et al. (2024)	Sycophancy evaluations and preference analyses	Agreement with user views can diverge from truth	Acquired recognition, longitudinal user learning, or performance of every current model
Blevins et al. (2026); P. Chen et al. (2026); Blevins (2026)	Computational dialogue completions; observational conversations and transcripts	Models and users can adapt linguistically; estimates differ in emphasis across settings	Causal effects of stable learned style on error detection or repair
Bao et al. (2025)	Controlled authorship-attribution study, N = 95	Style-based explanations can change reliance patterns without improving overall accuracy	Authorship explanations, not acquired familiarity with a recurring source style

References

- Anderl, C., Klein, S. H., Sarigül, B., Schneider, F. M., Han, J., Fiedler, P. L., & Utz, S. (2024). Conversational presentation mode increases credibility judgements during information search with ChatGPT. *Scientific Reports*, 14, Article 17127. <https://doi.org/10.1038/s41598-024-67829-6>
- Bao, C., Baumler, C., Daumé III, H., & Carpuat, M. (2025). Who's the author? How explanations impact user reliance in AI-assisted authorship attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 25312–25330). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-emnlp.1380>
- Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1), 2–11. <https://doi.org/10.1609/hcomp.v7i1.5285>
- Begg, I. M., Anas, A. P., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, 121(4), 446–458. <https://doi.org/10.1037/0096-3445.121.4.446>
- Berger, J., Burton, J. W., Hertwig, R., Kosch, T., Kurvers, R. H. J. M., Kurzenberger, B., Lazik, C., Onnasch, L., Rieger, T., Thoma, A. I., Wulff, D. U., & Herzog, S. M. (2026). Human learning is an understudied but promising lever for boosting human–AI synergy. *Proceedings of the National Academy of Sciences*, 123(39), Article e2536100123. <https://doi.org/10.1073/pnas.2536100123>
- Blevins, T. (2026). *Accommodation goes both ways: Studying linguistic convergence between humans and language models* (Version 1) [Preprint]. arXiv. <https://arxiv.org/abs/2605.29278v1>
- Blevins, T., Schmalwieser, S., & Roth, B. (2026). Do language models accommodate their users? A study of linguistic convergence. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 791–807). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.eacl-long.34>
- Bottoms, H. C., Eslick, A. N., & Marsh, E. J. (2010). Memory and the Moses illusion: Failures to detect contradictions with stored knowledge yield negative memorial consequences. *Memory*, 18(6), 670–678. <https://doi.org/10.1080/09658211.2010.501558>
- Brashier, N. M., Eliseev, E. D., & Marsh, E. J. (2020). An initial accuracy focus prevents illusory truth. *Cognition*, 194, 104054. <https://doi.org/10.1016/j.cognition.2019.104054>
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). 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), Article 188. <https://doi.org/10.1145/3449287>
- Cantor, A. D., & Marsh, E. J. (2017). Expertise effects in the Moses illusion: Detecting contradictions with stored knowledge. *Memory*, 25(2), 220–230. <https://doi.org/10.1080/09658211.2016.1152377>
- Cecil, J., Lermer, E., Hudecek, M. F. C., Sauer, J., & Gaube, S. (2024). Explainability does not mitigate the negative impact of incorrect AI advice in a personnel selection task. *Scientific Reports*, 14, Article 9736. <https://doi.org/10.1038/s41598-024-60220-5>
- Chen, A., Kim, S. S. Y., Franyutti-Cintron, A. N., Dharmasiri, A., Mukherjee, K., Russakovsky, O., & Fan, J. E. (2026). Presenting large language models as companions affects what mental capacities people attribute to them. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (pp. 1–30). Association for Computing Machinery. <https://doi.org/10.1145/3772318.3791049>
- Chen, J., Mishler, S., Hu, B., Li, N., & Proctor, R. W. (2018). The description–experience gap in the effect of warning reliability on user trust and performance in a phishing-detection context. *International Journal of Human–Computer Studies*, 119, 35–47. <https://doi.org/10.1016/j.ijhcs.2018.05.010>

- Chen, P., Guan, H., & Jeong, E. J. (2026). Who accommodates whom? Bidirectional linguistic accommodation and progressive interpersonal convergence in human–AI conversations. *Behavioral Sciences*, 16(5), Article 720. <https://doi.org/10.3390/bs16050720>
- de Sá Siqueira, M. A., Müller, B. C. N., & Bosse, T. (2024). When do we accept mistakes from chatbots? The impact of human-like communication on user experience in chatbots that make mistakes. *International Journal of Human–Computer Interaction*, 40(11), 2862–2872. <https://doi.org/10.1080/10447318.2023.2175158>
- Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, 14(2), 238–257. <https://doi.org/10.1177/1088868309352251>
- 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. <https://doi.org/10.1037/xge0000033>
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6), 697–718. [https://doi.org/10.1016/S1071-5819\(03\)00038-7](https://doi.org/10.1016/S1071-5819(03)00038-7)
- Eder, M., Lechner, C., & Jakesch, M. (2026). Effects of personality- and opinion-alignment in human-AI interaction (Version 3) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2511.10544>
- Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5), 993–1002. <https://doi.org/10.1037/xge0000098>
- Frisch, I., & Giulianelli, M. (2024). LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)* (pp. 102–111). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.personalize-1.9>
- Gamez-Djokic, M., & Waytz, A. (2026). The perceived whiteness of artificial intelligence. *Journal of Personality and Social Psychology*. Advance online publication. <https://doi.org/10.1037/pspi0000533>
- Guo, Z., Wu, Y., Hartline, J. D., & Hullman, J. (2024). A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 221–236). Association for Computing Machinery. <https://doi.org/10.1145/3630106.3658901>
- Hirst, G., Johnson, W., Li, A. J., & Richter, A. W. (2026). Workers shift their views and pay more when AI chatbots pander to their values. *Scientific Reports*. Advance online publication. <https://doi.org/10.1038/s41598-026-71409-1>
- Hoegen, R., Aneja, D., McDuff, D., & Czerwinski, M. (2019). An end-to-end conversational style matching agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (pp. 111–118). Association for Computing Machinery. <https://doi.org/10.1145/3308532.3329473>
- Jalbert, M., Schwarz, N., & Newman, E. (2020). Only half of what I'll tell you is true: Expecting to encounter falsehoods reduces illusory truth. *Journal of Applied Research in Memory and Cognition*, 9(4), 602–613. <https://doi.org/10.1016/j.jarmac.2020.08.010>
- Janoušková, A., Kocyan, J., Šimová, M., Uvířová, K., Zahradníčková, K., Vaculík, M., & Procházka, J. (2022). The effect of font readability on the Moses illusion: A replication study. *Consciousness and Cognition*, 99, 103284. <https://doi.org/10.1016/j.concog.2022.103284>
- Kamas, E. N., Reder, L. M., & Ayers, M. S. (1996). Partial matching in the Moses illusion: Response bias not sensitivity. *Memory & Cognition*, 24(6), 687–699. <https://doi.org/10.3758/BF03201094>

- Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, 32(1), 1–24. <https://doi.org/10.1006/jmla.1993.1001>
- Kim, A., Yang, M., & Zhang, J. (2023). When algorithms err: Differential impact of early vs. late errors on users' reliance on algorithms. *ACM Transactions on Computer-Human Interaction*, 30(1), Article 14. <https://doi.org/10.1145/3557889>
- Kim, S. S. Y., Liao, Q. V., Vorvoreanu, M., Ballard, S., & Vaughan, J. W. (2024). “I'm not sure, but...”: Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 822–835). Association for Computing Machinery. <https://doi.org/10.1145/3630106.3658941>
- Kim, S. S. Y., Vaughan, J. W., Liao, Q. V., Lombrozo, T., & Russakovsky, O. (2025). Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–19). Association for Computing Machinery. <https://doi.org/10.1145/3706598.3714020>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- 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. <https://doi.org/10.1016/j.obhdp.2018.12.005>
- Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, 48(2), 241–256. <https://doi.org/10.1518/001872006777724408>
- McBride, S. E., Rogers, W. A., & Fisk, A. D. (2014). Understanding human management of automation errors. *Theoretical Issues in Ergonomics Science*, 15(6), 545–577. <https://doi.org/10.1080/1463922X.2013.817625>
- Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3), 171–181. <https://doi.org/10.1037/1076-898X.7.3.171>
- Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, D. C. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, 43(2), 223–239. <https://doi.org/10.1006/ijhc.1995.1042>
- Osler, L. (2026). Hallucinating with AI: Distributed delusions and “AI psychosis”. *Philosophy & Technology*, 39, Article 30. <https://doi.org/10.1007/s13347-026-01034-3>
- Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12), 1865–1880. <https://doi.org/10.1037/xge0000465>
- Pilotti, M., Chodorow, M., & Schauss, F. (2009). Text familiarity, word frequency, and sentential constraints in error detection. *Perceptual and Motor Skills*, 109(3), 627–645. <https://doi.org/10.2466/pms.109.3.627-645>
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 237). Association for Computing Machinery. <https://doi.org/10.1145/3411764.3445315>
- Raees, M., & Papangelis, K. (2026). From trust to appropriate reliance: Measurement constructs in human-AI decision-making [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2604.23896>

- Rallapalli, S., Gallagher, S. K., Yurko, R., Brooks, T., Loughin, C., Sezgin, M., & Turri, V. (2026). Interpretable stylistic variation in human and LLM writing across genres, models, and decoding strategies [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2604.14111>
- Sanchez, J., Rogers, W. A., Fisk, A. D., & Rovira, E. (2014). Understanding reliance on automation: Effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science*, 15(2), 134–160. <https://doi.org/10.1080/1463922X.2011.611269>
- Schoeffer, J., Jakubik, J., Vössing, M., Kühl, N., & Satzger, G. (2025). AI reliance and decision quality: Fundamentals, interdependence, and the effects of interventions. *Journal of Artificial Intelligence Research*, 82, 471–501. <https://doi.org/10.1613/jair.1.15873>
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askill, A., Bowman, S. R., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=tvhaxkMKAn>
- Song, H., & Schwarz, N. (2008). Fluency and the detection of misleading questions: Low processing fluency attenuates the Moses illusion. *Social Cognition*, 26(6), 791–799. <https://doi.org/10.1521/soco.2008.26.6.791>
- Spillner, L., & Wenig, N. (2021). Talk to me on my level – Linguistic alignment for chatbots. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction* (pp. 1–12). Association for Computing Machinery. <https://doi.org/10.1145/3447526.3472050>
- Trippas, D., Kellen, D., Singmann, H., Pennycook, G., Koehler, D. J., Fugelsang, J. A., & Dubé, C. (2018). Characterizing belief bias in syllogistic reasoning: A hierarchical Bayesian meta-analysis of ROC data. *Psychonomic Bulletin & Review*, 25(6), 2141–2174. <https://doi.org/10.3758/s13423-018-1460-7>
- Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(1), 219–230. <https://doi.org/10.1037/0278-7393.33.1.219>
- Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12), 2293–2303. <https://doi.org/10.1038/s41562-024-02024-1>
- Ye, S., Attali, D., Ghazi, M., Cachia, A., Cassotti, M., & Borst, G. (2026). Systematic review and meta-analysis of the evidence for an illusory truth effect and its determinants. *Nature Communications*, 17, Article 3270. <https://doi.org/10.1038/s41467-026-70041-x>
- Yin, M., Vaughan, J. W., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper 279, pp. 1–12). Association for Computing Machinery. <https://doi.org/10.1145/3290605.3300509>