

Reciprocal Miscalibration in Personalized AI Correction

User model fidelity persuasive fit and feedback over time

Conceptual working paper

Version 0.1 | 16 October 2026

Luiza Scurtu

DOI: [10.5281/zenodo.23435387](https://doi.org/10.5281/zenodo.23435387)

Abstract

Persistent AI agents can condition advice and correction on disclosures, histories, and inferred preferences. Apparent personalization fit is insufficient evidence of calibration. A system may misrepresent its user while the user poorly distinguishes accurate from inaccurate advice; their coexistence is termed dual miscalibration. Reciprocal miscalibration is a narrower hypothesis: representation-conditioned outputs affect behaviour, that behaviour updates the operative representation, and the updated state sustains or worsens both representation error and miscalibrated human reliance. The framework separates user-model fidelity, contextual representation fit, perceived fit, persuasive efficacy, and advice accuracy. An accurate record of a false user belief remains distinct from endorsing that belief as true. Performative prediction, preference pollution, persuasion profiling, distributed cognition, narrative railroading, and human-AI feedback research supply direct antecedents. The contribution is an operational refinement of these problems for conversational correction, with falsifiable propositions and interventions on representation updating and correction. Feedback to the human about independently checked outcomes is distinguished from behaviour fed back into the system. A factorial study tests immediate effects; a longitudinal intervention tests reciprocal reinforcement. Appropriate reliance additionally depends on payoffs, checking costs, relative competence, and learning opportunities. Neither reciprocity nor the benefit of the proposed controls has been established by this conceptual synthesis.

Keywords: human-AI interaction; personalization; user models; appropriate reliance; persuasion; profiling; calibration; autonomy; epistemic authority

1. Introduction

An AI agent can appear to know a person in several ways. It may repeat disclosed facts, retrieve earlier interactions, infer preferences, adapt examples, anticipate constraints, or adopt a recurring way of addressing the user. These outputs may feel relevant, familiar, or understanding. They may also affect whether the user accepts a correction. Yet the subjective success of personalization does not establish that the system represents the person accurately, nor that the correction is substantively sound.

This paper develops that distinction into a framework for dual and reciprocal miscalibration. Human-AI reliance research distinguishes following advice, discriminating its quality, and achieving valuable final decisions (Lee & See, 2004; Guo et al., 2024; Schoeffer et al., 2025). User-modelling research examines how systems infer, expose, and update information about a user (Kay & Kummerfeld, 2012; Radlinski et al., 2022). Personalized language-model research demonstrates that histories and profiles can improve predictions or introduce new errors (Balepur et al., 2026; Park et al., 2026; Sun et al., 2026). Performative prediction and preference-pollution research already address how deployed systems alter their own later evidence (Perdomo et al., 2020; Adomavicius et al., 2022).

The paper's central proposition is:

In repeated human-AI correction, apparent fit alone cannot establish calibration. Increased receptivity need not indicate better discrimination, and increased persuasive efficacy need not indicate a more faithful representation of the user.

This proposition does not presume that personalization helps or harms. Accurate representations may improve relevance, reduce search costs, and make corrections easier to apply. Mistaken representations may be obvious and readily repaired. Equally, a sparse or inaccurate representation may generate an answer that happens to feel apt, while an accurate representation may support poor advice. The analytical problem is that acceptance, satisfaction, and feeling understood are compatible with several different underlying states.

The reciprocal element is a hypothesis about feedback over time. The system uses a representation to shape communication; communication affects judgement and behaviour; and later behaviour is used to update the representation. An inaccurate representation and poor human discrimination can coexist without this loop. Reciprocity requires evidence that the two directions are causally connected through updating, with subsequent consequences for both representation fidelity and human reliance. Predicting system-influenced behaviour is not by itself evidence of either miscalibration or harm.

The paper separates five constructs and specifies interventions on the operative representation. Conversational reinforcement, AI-shaped self-narratives, and deference to apparently understanding systems already have close theoretical and empirical antecedents (Osler, 2025a, 2025b, 2026; Dohnány et al., 2026; Glickman & Sharot, 2025). The contribution is their refinement into separately measurable states and causal tests of profile fidelity, human discrimination, updating, and effective correction. It does not claim discovery of conversational feedback or confirmation of a reciprocal effect.

2. Constructs and boundaries

2.1 Synthetic user representation

A **synthetic user representation** is retained or retrievable system state used to characterize a person for future interaction. It may include explicit profile fields, summaries, retrieved conversation history, behavioural features, inferred attributes, latent parameters, or combinations of these. The definition is functional: the representation must condition later selection, generation, ranking, advice, or correction.

Not every tailored response demonstrates a persistent representation. Adaptation to the current prompt may be transient. Nor does the presence of memory establish a coherent “model of the person.” A system may retrieve disconnected facts without integrating them, or may use a profile for one function but not another. Product-specific claims therefore require evidence about what state is retained, how it is retrieved, which outputs it conditions, and how updates propagate.

Four implementation channels must be distinguished: an explicit stored profile, retrieved interaction history, the current context window, and model-parameter updates. A conversation can change subsequent output through context alone, without retraining model weights. Retrieved prior dialogue falls within the functional definition above when it characterizes the person and conditions future interaction; an explicit profile is not required. Studies must identify the channel that carries the focal claim and log the state actually used. Freezing a profile while allowing influenced dialogue to re-enter through retrieval does not freeze the operative representation.

2.2 Five distinct variables

The framework separates five variables:

1. **User-model fidelity:** accuracy of the claims encoded in the operative representation against independently assessable facts or constraints at a specified time and context. Coverage of missing attributes is reported separately.
2. **Contextual representation fit:** adequacy of the attributes selected, covered, and weighted for the present task, including whether relevant information is current. It concerns what the representation makes available for this use, not whether the output feels apt.
3. **Perceived fit:** the person's subjective sense that an output is tailored, apt, responsive, or understanding.
4. **Persuasive efficacy:** the causal effect of a communication on acceptance, confidence, verification, or action relative to a specified comparator. Observed acceptance alone is not a causal efficacy estimate.
5. **Advice accuracy:** whether the substantive answer or correction is correct under an independently defensible standard.

User-model fidelity is deliberately narrower than a claim to capture the “real person.” For confirmatory work, it concerns checkable represented claims. Contextual fit concerns coverage, relevance, weighting, and current applicability; it overlaps with fidelity but is not interchangeable with it. For example, a profile can state budget and availability accurately yet omit a required accessibility constraint: high claim accuracy, inadequate coverage. Conversely, a task-adequate profile may contain an incorrect secondary transport cost while correctly covering the constraints that determine a feasible plan. Studies should score claim accuracy, coverage, and consequential errors separately rather than assert five independent latent variables. Perceived fit is a judgement about the output, persuasive efficacy is a causal behavioural contrast, and advice accuracy is evaluated against an independent answer standard.

2.3 Personal relevance is not familiarity

Profile conditioning is not sufficient to establish familiarity. A response may be personally relevant without being recognisable, tailored in content without having a stable style, or persuasive because it invokes a disclosed concern. This paper therefore does not treat representation fit as a species of style familiarity.

Representation-mediated familiarity is reserved for a narrower longitudinal phenomenon: repeated profile-conditioned interaction produces a recurring interpersonal or communicative pattern that the user acquires and recognises across new content. Demonstrating it requires exposure and a recognition measure. It may become one

pathway through which a user interprets personalized correction, but it is not assumed whenever a profile conditions output.

2.4 Calibration and correction

On the human side, this paper distinguishes sensitivity to advice correctness from the decision to accept, reject, or verify. On the system side, it evaluates whether retained claims and their use correspond to the person and task. “Calibration” is used in this relational sense; representation fidelity is not statistical calibration of predicted probabilities. Where systems make probabilistic claims, probability calibration requires its own measure. A person can discriminate advice well despite a poor profile, and an accurate profile can coexist with incorrect advice.

Appropriate reliance aims at expected decision value, given error prevalence, asymmetric consequences, verification costs, and relative human and AI competence. Discrimination is one component, not its complete normative definition (Guo et al., 2024; Schoeffer et al., 2025). A rational user lacking diagnostic evidence may accept some advice that later proves wrong. Studies should therefore predefine the payoff structure and compare decisions with a benchmark using the information available at the time, rather than classify every incorrect outcome as irrational reliance.

2.5 Representing a belief and endorsing its truth

Suppose a user wrongly believes that a colleague is sabotaging them. A profile stating “the user believes the colleague is sabotaging them” may accurately describe the user. Advice stating “the colleague is sabotaging you” makes a separate claim about the world and can be false. Coding the first as low user-model fidelity merely because the underlying belief is false would collapse the constructs. Studies must distinguish attributed beliefs, externally checkable facts, and interpretive self-narratives.

Consequently, accurate personalization can reinforce a mistaken belief through selective affirmation. Sycophancy research documents model agreement with user views at the expense of truth in the tested systems (Sharma et al., 2024). Osler (2026) explicitly considers both AI-originating falsehoods and reinforcement of user-originating beliefs. These pathways need not involve an inaccurate user representation. Reciprocal miscalibration, as defined here, is one proposed failure pattern within this broader field. It is not a necessary explanation of every harmful conversational feedback loop. Intentional fiction, aspirational self-description, and informed goal revision also require different evaluation standards.

3. Review method

This is a structured narrative review and conceptual synthesis, not a systematic review. The review was conducted across scholarly proceedings, publisher and DOI pages, arXiv, institutional repositories, official European data-protection sources, and public scholarly indexes. Targeted primary-source checks covered performative prediction, preference pollution, persuasion profiles, decision-theoretic reliance, and dynamic profile updating. Search strands covered profiling, personalization failures, influence, endogenous feedback, scrutable user models, and appropriate reliance. Computational benchmarks are used to establish technical possibility or failure modes, not human behavioural effects. Governance sources are used for institutional or normative claims, not causal evidence.

Within this bounded narrative search, however, no located study jointly manipulated user-model fidelity and advice correctness while tracing system-induced behaviour, its reuse in operative profile updating, and subsequent human discrimination over repeated conversational correction. This is a retrieval limit, not a claim that no prior study exists. The contribution rests on construct separation and testable application to this configuration, not on an absence claim.

4. What the evidence establishes

4.1 Histories can support profiles and predictions

Interaction histories can be converted into profiles that improve task performance. Park et al. (2026) retrieved user history and generated context-aware profiles for personalized persuasiveness prediction. Their reported improvement establishes predictive utility within the benchmark. It does not show that profile-conditioned messages causally persuade users, that the profiles faithfully represent them, or that persuasion improves their decisions.

Two 2026 systems make dynamic updating technically concrete. Personalized Agents from Human Feedback uses pre-action clarification, preference-conditioned action, and post-action feedback to update explicit per-user memory, evaluating preference learning and adaptation to persona shifts (Liang et al., 2026). PersonalAgent continually refines a unified user profile from dialogue and includes human evaluation of preference capture (Zhang et al., 2026). These are technical demonstrations of continual personalization. Neither independently manipulates profile error, advice correctness, system-induced behaviour, reuse of that behaviour, and later human discrimination or decision loss. They narrow the technical novelty claim without establishing reciprocal miscalibration.

The distinction matters because prediction targets observed behaviour, not necessarily antecedent preference or welfare. A system may predict what a person will endorse without correctly representing why, whether the endorsement is stable, or whether the predicted action serves the person's goals.

4.2 Personalization can introduce or conceal errors

Balepur et al. (2026) compared personalized deep-research systems using synthetic evaluation and real users. Real-user interviews revealed nuanced personalization errors that automated evaluation did not capture. The result supports the need for human-grounded evaluation and warns against treating synthetic profile benchmarks as sufficient. It does not establish how those errors affect trust or correction.

Sun et al. (2026) demonstrated personalization-induced factual distortion in benchmark experiments: user histories could pull model reasoning toward personalized content rather than factual correctness. This establishes a technical failure mode in which personalization and factual accuracy diverge. It does not establish that people find the distorted answers more credible, or that persistent interaction amplifies the effect.

Together these studies show that technically successful use of history can coexist with errors invisible to common evaluation methods. They do not show that felt fit rises as fidelity falls.

4.3 Behavioural feedback can contaminate later evidence

Performative prediction formalizes settings in which deploying a prediction changes the outcomes or distribution against which it is subsequently assessed (Perdomo et al., 2020). It is a direct antecedent to the present identification problem: a conversational agent may partly produce the behaviour on which its next user model is based. The present paper applies this problem to a retained representation and human correction; it does not introduce endogenous model–user feedback as a new mechanism.

Preference pollution is particularly close. Adomavicius et al. (2022) examine how recommendations can shift subsequent preference ratings that are then reused as ground truth. Their account distinguishes this contamination from selective exposure and non-representative observations. Chaney et al. (2018) additionally demonstrate in simulation that recommender feedback can increase homogeneity and reduce utility. These results motivate, but

do not demonstrate, corresponding conversational dynamics. Dialogue may change considered preferences legitimately; system influence alone is insufficient to classify the resulting state as inaccurate or harmful.

4.4 Personalized persuasion and perceived fit

Persuasion profiling already adapts strategies using explicit susceptibility measures and implicit behavioural responses (Kaptein et al., 2015). A representation can therefore be useful for influencing behaviour without being an adequate general account of the person. The distinction advanced here is between profile accuracy, contextual adequacy, subjective recognition or fit, causal persuasive effect, and feedback into later profile state. It is a refinement against this established literature, not a claim that personalization and persuasion were previously disconnected.

Value-congruent chatbot framing can increase endorsement and willingness to pay (Hirst et al., 2026). The reported pathways differ: a more compelling case contributes to endorsement and commercial engagement, while feeling understood additionally contributes to commercial engagement. This does not establish that feeling understood mediates endorsement, nor that an inferred persistent profile is accurate. Perceived fit is a candidate influence pathway requiring its own causal test.

4.5 Inspection does not guarantee effective correction

Scrutable user-model research has long argued that people should be able to inspect and control representations used for personalization (Kay & Kummerfeld, 2012). Natural-language user profiles have been proposed as a transparent and interrogable interface for recommendation (Radlinski et al., 2022). These contributions establish design principles and technical approaches, not that inspection reliably restores accuracy or autonomy.

Warshaw et al. (2015) found in qualitative research on hyper-personal analytics that some participants treated algorithmically generated profiles as authoritative and felt poorly qualified to modify them even when editing was possible. The study was small and predates current persistent language-model agents, but it shows why nominal editability and practical contestability should be measured separately.

European data-protection guidance treats profiling and automated decision-making as governance concerns (European Data Protection Board, 2018). Wachter and Mittelstadt (2019) argue that protections are materially weaker or less direct for inferred data and propose a right to reasonable inferences. The latter is a scholarly legal proposal, not a settled general entitlement. “Information sovereignty” is therefore used here as a normative umbrella for effective knowledge and control, not as the name of a universally established legal right.

4.6 Conversational feedback and narrative authority

The closest conceptual overlap is already conversational. Osler (2026) treats AI as both a cognitive artifact and an apparently responsive social other, explaining how dialogue can scaffold and sustain mistaken beliefs. The argument extends beyond clinical delusions to everyday interpretations and self-narratives, including personalized interaction and memory. It therefore cannot be set aside as only clinical or as lacking persistent representations. The article offers a philosophical account, not a controlled estimate of causal effects or prevalence.

Osler’s narrative railroading account (2025b) connects algorithmic representations, tailored information, self-narratives, and behaviour. Her chapter manuscript (2025a) examines chatbot-assisted self-understanding and narrative construction. Byrne (2025) supplies an adjacent account of narrative deference; Clowes (2020) examines digital representations of persons, and Telakivi (2026) distinguishes forms of AI-mediated remembering. These sources establish substantial antecedents for questions about representation and authority. They do not supply the joint operational measures or randomized update interventions proposed here.

Dohnány et al. (2026) likewise model bidirectional chatbot–user reinforcement and include simulations in which both interlocutors are AI agents. These are mechanism demonstrations under simulated assumptions, not evidence that real human participants developed the corresponding changes. In contrast, Glickman and Sharot (2025) provide human experiments showing that interaction with biased AI can alter later human judgements, while accurate AI can improve them. Their studies make feedback and human learning direct empirical antecedents, without testing the present conjunction of an independently scored personal representation and human advice discrimination.

Real-world conversational evidence also requires care. Sharma et al. (2026) classify potential disempowerment patterns in LLM conversations. Those classifications are not verified changes in a user’s values, clinical states, or longitudinal outcomes. Taken together, these works require a modest contribution claim: the present framework specifies which representation claims are wrong, which human reliance outcomes fail, and which update intervention would identify their connection. The existence of conversational reinforcement itself is not the proposed novelty.

4.7 Outcome feedback and opportunities to learn

Berger et al. (2026) reanalysed the 74 studies in Vaccaro et al. (2024): only ten provided outcome feedback, and none randomized its availability. The feedback-explanation combination had a positive synergy estimate, but its credible interval included zero; explanations without feedback had a negative estimate. These cross-study associations motivate causal tests; they do not establish that feedback improves collaboration.

Human-AI synergy compares joint performance with the stronger partner alone; it is distinct from advice discrimination and retained unaided learning. The framework therefore treats opportunities to learn as part of the task environment. AI-generated reassurance or explanation cannot serve as independent verification of the answer. A lack of accessible outcome information also limits what a user could rationally learn about either their own competence or the agent’s.

5. Dual states and the reciprocal feedback hypothesis

5.1 The representation - influence loop

Figure 1 depicts a possible representation - influence loop. Disclosures and observed behaviour contribute to retained user state, which conditions advice or correction. Human responses generate new traces that may update that state. The presence of a loop is not sufficient for miscalibration: each relevant departure from a fidelity or decision standard must also be measured.

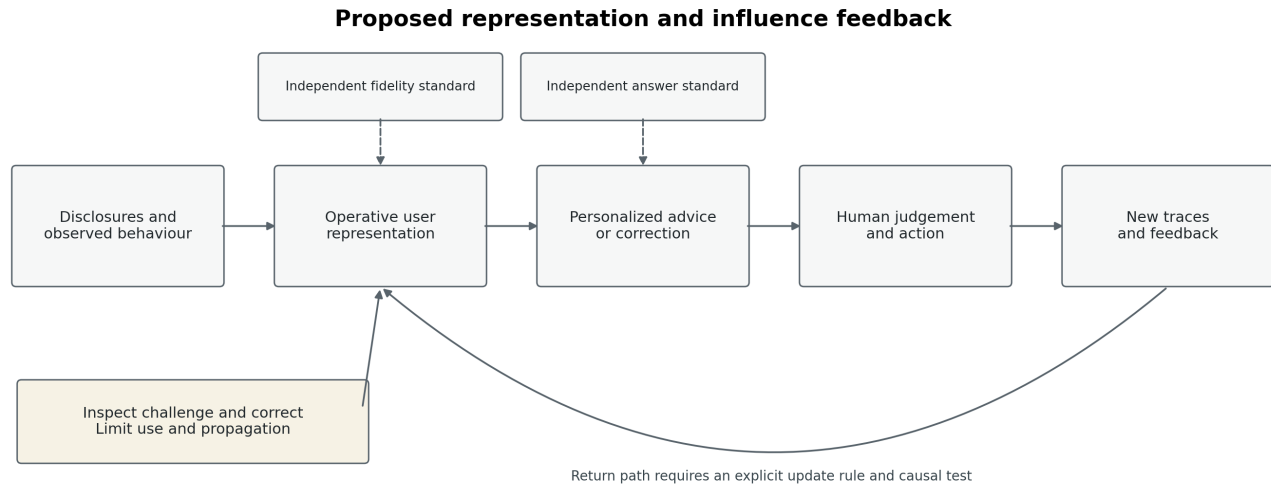


Figure 1. Proposed representation and influence feedback. Independent standards evaluate fidelity and advice accuracy. The return path depicts behaviour reused by the system; independent outcome feedback to the human is a separate channel described in Section 5.5. Neither path is an observed result of this paper.

The loop contains two external standards. **Fidelity** requires comparison between the operative representation and the person's current task-relevant constraints. **Accuracy** requires comparison between advice and an independent answer standard. Perceived fit and acceptance occur inside the loop and cannot validate either standard by themselves.

The identification problem is that later behaviour is not independent evidence of antecedent preference when prior outputs helped shape it. Reciprocal miscalibration is the narrower hypothesis that a mistaken operative representation changes responses in ways learned as support for that representation, with the updated state sustaining or worsening both representation error and miscalibrated human reliance. The process can maintain an error without a monotonic decline in every measure. A profile feedback effect without a measured human reliance failure is insufficient for this stronger claim. Cross-sectional coexistence, contemporaneous interactions, or rising predictive performance cannot establish it.

5.2 Four static configurations

Table 1 crosses representation fidelity with behavioural discrimination. These are descriptive dimensions, not a complete normative assessment of reliance. None of the four cells establishes reciprocity.

User model fidelity	Advice discrimination	Static interpretation
Higher fidelity	Better discrimination	Aligned measured dimensions. Neither optimal reliance nor reciprocal feedback follows from this cell alone.
Higher fidelity	Poorer discrimination	Accurate represented claims with poor discrimination of advice quality.
Lower fidelity	Better discrimination	Representation error with preserved advice discrimination. The person need not have detected the profile error.
Lower fidelity	Poorer discrimination	Dual miscalibration. Coexistence of the two failures; no causal reinforcement is implied.

Table 1. Static configurations of representation fidelity and advice discrimination. Higher and lower values require task-specific, preregistered standards; these are not natural binary categories.

Dual miscalibration names the static combination of lower representation fidelity and poorer advice discrimination. It does not imply that one caused the other. Higher discrimination can coexist with an inappropriate decision threshold, and poor discrimination does not make every act of reliance irrational. The labels are task- and time-specific; longitudinal evidence is required to infer movement or reinforcement between states.

5.3 Apparent fit is not validation

Perceived fit can be accurate, coincidental, strategically produced, or generated by repetition. Persuasive efficacy can reflect useful communication or compliance. Advice accuracy can be high despite a poor representation when the correct answer does not depend on personal information. Conversely, a faithful representation can be used to deliver bad advice more effectively.

The evaluative rule is therefore:

Do not infer user-model fidelity or advice accuracy from perceived fit, satisfaction, acceptance, or prediction of behaviour alone.

The rule applies to two distinct validation problems. Human acceptance alone cannot validate the answer; predictive success on behaviour the system influenced cannot validate its representation of antecedent user states. Both cautions have established antecedents in reliance and endogenous-feedback research.

5.4 Representation-mediated familiarity as an emergent pathway

Repeated personalized interaction may generate a recognisable interpersonal pattern. If users learn and recognise that pattern across new content, representation-mediated familiarity may become a second-order cue in reliance. That pathway is plausible but conditional. It requires evidence of acquisition and recognition, not merely profile use or self-relevance.

This construct links, but does not merge, reciprocal calibration with acquired-style calibration. A user might recognise how an agent typically personalizes corrections and use departures as a signal. Equally, the familiar pattern might increase acceptance without improving discrimination. Those possibilities require a longitudinal recognition design and should not be inferred from a one-session personalization effect.

5.5 Two feedback channels

Behavioural feedback to the system supplies traces for a later representation. Outcome feedback to the human supplies independently warranted information about correctness or consequences. These channels can operate together or separately. A system can repeatedly learn from acceptance while the person receives no reliable evidence that the accepted answer was wrong. Conversely, independent feedback may help the person challenge a mistaken inference and may interrupt its reuse. The latter is a testable boundary hypothesis, not an established remedy.

Specify the source, timing, reliability, and availability of human outcome feedback, and whether it evaluates the initial human answer, the AI answer, or the final decision. For system updating, specify which traces are reused, their exposure provenance, and the state they modify. Calling both processes “feedback” without those distinctions obscures the causal question.

6. What the framework contributes

The framework refines established personalization, narrative, persuasion, and conversational-feedback accounts into a research design for correction. Its contribution has three parts.

First, it separates claim accuracy and contextual coverage from perceived fit, persuasive effects, and substantive advice accuracy. This permits divergence to be measured without treating satisfaction or acceptance as sufficient evidence of success.

Second, it distinguishes dual miscalibration from a reciprocal process. A factorial design can identify immediate effects and statistical interactions; it cannot identify a feedback mechanism that it does not manipulate or observe over time.

Third, it specifies a test of whether influenced behaviour is reused as evidence for an inaccurate representation, and whether correcting or interrupting that update changes later representation and human outcomes. This is an operational and intervention-focused contribution within existing conversational-feedback research, not the discovery of an endogenous loop. It also distinguishes opportunities for human learning from the system's reuse of behaviour.

Five falsifiable propositions organise the proposed research. They are conditional predictions, not findings or claims that all personalized interaction follows these paths.

P1 Perceived-fit influence. For a profile-conditioned framing independently validated to raise felt relevance without adding task evidence, the proposed pathway predicts greater acceptance of correct and incorrect advice without improved discrimination. The test is a criterion or acceptance shift in the full randomized framing contrast, not an analysis restricted to participants reporting greater fit. A precise null weakens this pathway; a sensitivity gain requires an additional explanation.

P2 Effective correction. Where a checkable profile error is causally used in later outputs, verified editing of the operative state should reduce recurrence of that error relative to read-only inspection. A changed visible summary without changed downstream state does not satisfy this proposition. No recurrence reduction despite verified updating weakens the proposed control benefit.

P3 Reciprocal reinforcement. With fixed antecedent constraints and a preregistered updating rule, feeding system-influenced behaviour back as unqualified user evidence should sustain more profile error and greater human decision loss, or poorer advice discrimination, than blocking that feedback or retaining its provenance. Attribution requires an influence on behaviour, update-dependent later profile error, and a corresponding human reliance failure. A profile-only effect or absence of either causal link weakens the stronger reciprocal hypothesis.

P4 Goal revision boundary. Reinforcement attributed to a mistaken representation should diminish when updates distinguish AI-exposed behaviour from independently reaffirmed constraints or goals. If later changes instead reflect informed, independently endorsed goal revision and improve relevant outcomes, the process should be classified as adaptive change rather than miscalibration. This competing explanation is tested with timing, independent reaffirmation, and outcome measures.

P5 Independent feedback boundary. Where participants have sufficient task knowledge and receive timely, reliable outcome information, feedback is predicted to reduce the persistence of the P3 pattern relative to otherwise matched interaction without that information. The prediction concerns the measured profile-error and human-outcome trajectory, not trust alone. No reduction in a sufficiently precise randomized contrast weakens this proposed interruption mechanism. Unreliable feedback, delayed outcomes, or lack of knowledge needed to interpret them are boundary conditions.

The propositions separate immediate influence (P1), effective profile correction (P2), reciprocal reinforcement and legitimate revision (P3–P4), and an independent-feedback boundary (P5). P1 and P2 can hold while reciprocity is absent. Direction, magnitude, prevalence, and generalisation remain empirical questions.

7. Future empirical test

7.1 Confirmatory study using checkable constraints

The smallest useful factorial study should use checkable constraints rather than subjective preferences as ground truth. Participants could complete a planning or allocation task with documented budget, availability, exclusions, or responsibilities. The focal constraints must matter to the recommendation and remain fixed during the confirmatory phase. This study tests immediate personalization effects and dual-state configurations, not reciprocity.

Construct an operative profile that is accurate, selectively incomplete, or contains a consequential mistaken constraint, and cross this assignment with advice correctness. Incomplete profiles must be distinguished from false claims. Use independently authored, prevalidated advice variants so that profile fidelity and correctness can vary separately; an unconstrained generator could otherwise collapse the factors. Record whether a defect concerns a user constraint or substantive task reasoning, and match error severity and detectability where possible. Hold style, confidence, clarity, competence history, and source label constant.

Representation control can be added in a separately powered stage: opaque, inspectable but read-only, or editable with verification that the operative state changes. Test later propagation of the edit. To test P1 specifically, also randomize a content-matched personal-relevance framing against neutral framing; profile fidelity alone does not identify the causal effect of perceived fit.

Preregister a small primary endpoint set: advice-discrimination sensitivity and final constraint-respecting accuracy, with recurring profile errors primary in the control stage. Report perceived fit, acceptance, verification, defect identification, and successful profile correction separately. Use participant and item effects, specified interaction contrasts, simulation-based power, multiplicity control, and intention-to-treat analysis. Treat perceived fit as an outcome. A statistical indirect effect does not establish causal mediation: mediator–outcome confounding and post-treatment confounding require explicit assumptions, sensitivity analysis, or a separate intervention on the proposed pathway.

Record an unaided baseline and the information participants can use to assess their own and the AI's competence. Specify one independently checked outcome-feedback regime and hold its content and timing constant across the profile and framing assignments. Include feedback-free assessment probes before revealing answers so that immediate instruction is not scored as unaided learning. Analyse initial-to-final decision transitions as well as aggregate accuracy.

7.2 Preferences as a separate exploratory construct

Values and preferences should not be treated as equivalent to checkable constraints. They may change, conflict, or become clearer through deliberation. A separate exploratory phase can examine participant-endorsed preferences, but it must specify timing, context, confidence, and the procedure for resolving disagreement between prior disclosure, later endorsement, and observed behaviour. Self-report is evidence, not an infallible ground truth.

7.3 Longitudinal test of reciprocity

To test P3, record checkable constraints, randomize an accurate or mistaken starting profile and matched framing, and observe subsequent choices. Cross this assignment with a system-update intervention: block reuse of AI-influenced behavioural traces, permit a preregistered behaviour-based update, or retain provenance so that those traces are not treated as independent evidence. Log the complete operative state and specify whether the

intervention changes a profile, retrieval, context, or parameters. Prevent an ostensibly blocked trace from returning through another channel. Keep human verification opportunities, outcome-feedback schedules, and task exposure comparable across update arms. Estimate whether initial error affects behaviour and whether its reuse sustains later profile error together with poorer human discrimination or decision value. Both dimensions and both causal links are required; temporal association or cross-lagged correlation alone is insufficient.

For P4, a separate preference phase should distinguish independently reaffirmed goals from choices made under the agent's framing. Allow participants to revise goals explicitly; do not presume every shift is contamination. Debrief low-stakes framing manipulations, avoid consequential deception, and obtain any required ethics approval before recruitment. Preregister update rules, longitudinal estimands, attrition handling, stopping rules, exclusions, and smallest effects of interest. Archive materials, state trajectories, generated outputs, and de-identified data and code where consent and privacy permit.

A reciprocal mechanism is supported only by the prespecified chain and its update-dependent consequence, not by a generic interaction term. Frozen and provenance-aware updates provide discriminating counterfactuals. Beneficial learning, ordinary preference change, selective exposure, and a purely behavioural persuasion effect remain alternatives when the full chain is absent.

7.4 A separate test of the human learning boundary

Test P5 in a separately powered longitudinal experiment by randomizing independently checked outcome feedback while keeping style, task accuracy, and explanation availability fixed. Retain a prespecified system-update contrast if the target is feedback's interruption of reciprocity; a feedback main effect alone cannot establish that interaction. Score the P3 chain, learning across preregistered trial blocks, and later unaided transfer to new items. Assess retention after a prespecified delay, with no AI assistance or answer feedback during that assessment. If explanation availability is itself a target, vary it in a separately powered factorial extension rather than infer its effect from the feedback comparison. This module is not an added requirement within the immediate-effects study.

8. Human epistemic authority and control

When a representation shapes how an agent corrects a person, meaningful control concerns more than access to stored facts. Five properties should be evaluated separately:

- **provenance:** which disclosures, observations, external sources, and inferences contributed to the representation;
- **inspectability:** whether the person can see a meaningful account of the operative representation and its use;
- **contestability and update efficacy:** whether a challenge changes the state that actually conditions later output;
- **scope and propagation:** whether correction reaches retained copies, retrieval indexes, downstream uses, or other agents; and
- **delegation authority:** who may reuse the representation, for which tasks, and under whose authorization.

These are provisional evaluation and governance proposals. The present synthesis does not establish that any particular interface or control improves decision quality, or that every listed property is a legal requirement in every jurisdiction.

An editable profile can coexist with hidden latent state. A corrected interface can coexist with an uncorrected downstream copy. Disclosure can coexist with practical inability to contest an inference. Control is therefore an empirical system property, not a conclusion that follows from the presence of a settings screen.

Effective user control and factual correction are separate outcomes. A requested edit may propagate successfully while preserving a false belief; that demonstrates control, not truth. Conversely, disagreement with an interpretation does not by itself authorize treating the person as mistaken about every aspect of their identity. Audit edit propagation, factual accuracy, and endorsement of preferences separately. The narrative-authority literature makes this distinction consequential without supplying a single external truth standard for every self-description (Byrne, 2025; Osler, 2025a, 2025b).

Personal and professional representation should not be treated as a clean binary. Personal inference may implicate identity, intimacy, or vulnerability. Professional misclassification can affect livelihood and reputation. The relevant dimensions are sensitivity, persistence, contextual integrity, downstream reach, reversibility, contestability, and stakes.

The concept of human epistemic authority used here is limited: the person's practical ability to understand why a correction is being framed in a particular way, challenge the representation that shaped it, and prevent a disputed inference from silently governing later interactions. It does not imply that a person's preferred self-description automatically determines every factual or institutional question.

9. Material weaknesses and boundary conditions

1. **The specified reciprocal conjunction remains untested.** Static dual miscalibration and immediate interaction effects are insufficient to establish feedback or reinforcement.
2. **Fidelity is attribute-specific.** A representation may be accurate for scheduling constraints and wrong about values. Studies must specify the represented attribute, time, context, and standard.
3. **The person is not static ground truth.** Preferences, identities, and goals can change or conflict. Confirmatory claims should rely on checkable constraints; broader constructs require plural evidence.
4. **Perceived fit is easy to overinterpret.** Feeling understood, recognising details, liking the style, and receiving useful adaptation are distinct outcomes.
5. **Advice may not depend on the profile.** If the correct answer is invariant across users, representation error may have no practical effect. Tasks must make the focal constraint consequential.
6. **Inspectability may be performative.** A readable summary may omit hidden features or fail to control the state used downstream.
7. **Conversational reinforcement has direct antecedents.** Distributed-delusion accounts, narrative railroading, chatbot simulations, and human–AI feedback experiments materially constrain novelty. The untested claim here is the specified conjunction of representation error, human reliance failure, and update-dependent reinforcement.
8. **Persistent-agent implementations differ.** Claims about named systems require product-specific audits of memory, retrieval, inference, sharing, and correction controls.
9. **Governance scope can overwhelm the mechanism.** Privacy, data protection, autonomy, delegation, and interoperability are related but not resolved by this framework.
10. **Personalization can be beneficial.** The framework should not be read as an argument against retained context; it specifies evidence needed to distinguish useful adaptation from persuasive misfit.

11. **Accurate profiles can support false endorsement.** The proposed reciprocal pattern does not exhaust sycophancy, narrative dependence, or other harms. Clinical examples do not establish diagnoses, prevalence, or causation in ordinary use.
12. **Learning conditions can alter the trajectory.** Independent outcome information and relative competence must be measured; absence of feedback does not establish a stable human inability to discriminate.

These limitations preclude claims that reciprocal miscalibration has been demonstrated or that particular governance controls are proven remedies. They are compatible with publication of a bounded conceptual working paper and proposed research programme.

10. Conclusion

Persistent AI correction involves how people use advice and how retained user representations guide that advice. Apparent personalization fit cannot establish representation fidelity, substantive accuracy, or optimal reliance.

Dual miscalibration describes coexisting representation error and poor human discrimination. Reciprocal miscalibration requires evidence of a narrower update-dependent causal chain affecting both dimensions. Conversational reinforcement and AI-shaped self-narratives already have close antecedents. The contribution is to distinguish representation fidelity from belief truth and perceived fit, and to specify interventions on updating, provenance, and operative correction.

The factorial study tests immediate effects; the longitudinal intervention tests reciprocity. Independent outcome feedback to the human is distinguished from behavioural evidence reused by the system, with learning and unaided transfer evaluated separately. The framework leaves beneficial adaptation and informed goal revision possible. Increasing acceptance, behavioural predictability, or user control alone cannot establish increasing understanding or substantive accuracy.

Acknowledgements and provenance

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

Appendix A. Concise evidence register

Source	Evidence type	What it supports	What it does not establish
Park et al. (2026)	History-derived profiles for persuasiveness prediction	Profiles can improve prediction of persuasive outcomes	Causal persuasion, profile fidelity, or decision quality
Balepur et al. (2026)	Personalized deep-research evaluation with real users	Real users identify nuanced errors missed by synthetic evaluation	Longitudinal influence or correction of a persistent profile
Sun et al. (2026)	Personalized-language-model benchmarks	User histories can distort factual reasoning toward personalized content	Human credibility, reliance, or behavioural effects
Chaney et al. (2018)	Recommender-system simulations	System-shaped behaviour can feed back as training evidence and reduce utility	Equivalent dynamics in conversational agents

Source	Evidence type	What it supports	What it does not establish
Hirst et al. (2026)	Preregistered value-congruent chatbot experiments	Congruent framing can increase endorsement and commercial engagement; feeling understood is a reported pathway for commercial engagement	Inferred profiles, longitudinal familiarity, or calibration
Kay and Kummerfeld (2012)	Scrutable user-model principles and systems	Inspectable and controllable user representations are technically tractable	That inspection reliably corrects operative state or restores autonomy
Radlinski et al. (2022)	Natural-language profile approach	Profiles can be expressed in a potentially interrogable form	Behavioural benefit or complete representation control
Warshaw et al. (2015)	Qualitative study of hyper-personal profiles	Users may defer to algorithmic profiles despite nominal editability	Current persistent-agent behaviour or causal effects
European Data Protection Board (2018)	Official profiling guidance	Profiling and automated decision-making raise data-governance duties	A comprehensive right over every inference
Wachter and Mittelstadt (2019)	Legal analysis and normative proposal	Existing protections for inferred data may be limited or indirect	Settled universal law or behavioural effects
Yin et al. (2019)	Randomised studies of accuracy and reliance	Experienced accuracy changes behavioural reliance	User-representation fidelity or personalization effects
Perdomo et al. (2020)	Formal performative-prediction framework	Deployment can alter the target distribution used for later evaluation	Reciprocal conversational miscalibration or an inaccurate user profile
Adomavicius et al. (2022)	Synthesis of preference non-representativeness and pollution	System recommendations can contaminate ratings later used as ground truth	That every preference change is harmful or that dialogue has the same effect
Kaptein et al. (2015)	Persuasion profiling and three case studies	Persuasive strategies can adapt using explicit and implicit user responses	Advice accuracy, broad profile fidelity, or human error discrimination
Guo et al. (2024); Schoeffer et al. (2025)	Formal reliance and decision-quality frameworks	Following advice, discriminating signals, and decision value must be distinguished	The present feedback hypothesis or a universal optimal threshold
Osler (2025a, 2025b, 2026); Byrne (2025)	Philosophical accounts and chapter manuscript	Conversational reinforcement, narrative shaping, and deference are close antecedents	Causal effects, prevalence, or the present jointly measured profile-reliance process
Clowes (2020); Telakivi (2026)	Conceptual analyses of person representations and AI memory	Representation and memory can shape agency and self-understanding	Measured efficacy of the proposed correction controls
Dohnány et al. (2026)	Perspective including AI-agent simulations	A bidirectional conversational mechanism under simulated assumptions	Human clinical outcomes or independently scored user-model fidelity
Glickman and Sharot (2025)	Human experiments with biased and accurate AI	Interaction can change later human judgement in either direction	The specified personal-profile error and reliance feedback conjunction
Sharma et al. (2024, 2026)	Model evaluations; observational conversation classifications	Truth and agreement can diverge; potential disempowerment patterns can be coded	That accurate profiles prevent false endorsement or that coded risks are verified harms

Source	Evidence type	What it supports	What it does not establish
Berger et al. (2026); Vaccaro et al. (2024)	Reanalysis of collaboration studies; original meta-analysis	Learning opportunities deserve explicit design and reporting	A randomized feedback effect, retained learning, or proof of the P5 boundary
Liang et al. (2026); Zhang et al. (2026)	Continual-personalization systems with dynamic per-user memory or profiles	Interaction feedback can update user-specific state; PersonalAgent includes human evaluation	System-shaped feedback reuse, independently manipulated profile error and advice correctness, or later human reliance failure

References

- Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2022). Recommender systems, ground truth, and preference pollution. *AI Magazine*, 43(2), 177–189. <https://doi.org/10.1002/aaai.12055>
- Balepur, N., Hamada, M., Kishore, V., Feldman, S., Singh, A., Siangliulue, P., Chang, J. C., Choi, E., Boyd-Graber, J. L., & Naik, A. (2026). Language models don't know what you want: Evaluating personalization in deep research needs real users. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15910–15944). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.acl-long.723>
- 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>
- Byrne, E. A. (2025). Narrative deference. *Topoi*, 44(2), 405–417. <https://doi.org/10.1007/s11245-024-10105-z>
- Chaney, A. J. B., Stewart, B. M., & Engelhardt, B. E. (2018). How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 224–232). Association for Computing Machinery. <https://doi.org/10.1145/3240323.3240370>
- Clowes, R. W. (2020). The internet extended person: Exoself or doppelganger? *Limite*, 15, Article 22. <https://revistalimite.uta.cl/index.php/limite/article/view/234>
- Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M., & Nour, M. M. (2026). Technological folie à deux: Feedback loops between AI chatbots and mental health. *Nature Mental Health*, 4(3), 336–345. <https://doi.org/10.1038/s44220-026-00595-8>
- European Data Protection Board. (2018, May 25). Automated decision-making and profiling [Endorsement of the Article 29 Working Party guidelines WP251 rev.01]. https://www.edpb.europa.eu/documents/guideline/automated-decision-making-and-profiling_en
- Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9(2), 345–359. <https://doi.org/10.1038/s41562-024-02077-2>
- 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>
- Kaptein, M., Markopoulos, P., de Ruyter, B., & Aarts, E. (2015). Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies*, 77, 38–51. <https://doi.org/10.1016/j.ijhcs.2015.01.004>
- Kay, J., & Kummerfeld, B. (2012). Creating personalized systems that people can scrutinize and control: Drivers, principles and experience. *ACM Transactions on Interactive Intelligent Systems*, 2(4), Article 24. <https://doi.org/10.1145/2395123.2395129>
- 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
- Liang, K., Kruk, J., Qian, S., Yang, X., Bi, S., Yao, Y., Nie, S., Zhang, M., Liu, L., Fernández Fisac, J., Zhou, S., & Hosseini, S. (2026). *Learning personalized agents from human feedback* (Version 1) [Preprint]. arXiv. <https://arxiv.org/abs/2602.16173v1>
- Osler, L. (2025a). Knowing oneself through and with AI: From self-tracking to chatbots (Version 1) [Chapter manuscript]. arXiv. <https://arxiv.org/abs/2512.03682v1>
- Osler, L. (2025b). Narrative railroading. *Topoi*, 44(2), 379–392. <https://doi.org/10.1007/s11245-024-10146-4>
- 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>
- Park, S., Park, Y., Lim, J., & Jo, Y. (2026). Learning to retrieve user history and generate user profiles for personalized persuasiveness prediction. In *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 17338–17359). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.findings-acl.858>
- Perdomo, J., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning (Vol. 119, pp. 7599–7609)*. *Proceedings of Machine Learning Research*. <https://proceedings.mlr.press/v119/perdomo20a.html>

- Radlinski, F., Balog, K., Diaz, F., Dixon, L., & Wedin, B. (2022). On natural language user profiles for transparent and scrutable recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2863–2874). Association for Computing Machinery. <https://doi.org/10.1145/3477495.3531873>
- Schoeffer, J., Jakubik, J., Vössing, M., Kühn, 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., McCain, M., Douglas, R., & Duvenaud, D. (2026). Who’s in charge? Disempowerment patterns in real-world LLM usage (Version 1) [Preprint]. arXiv. <https://arxiv.org/abs/2601.19062v1>
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Asbell, 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>
- Sun, Z., Zhan, Y., Shen, C., Yu, W., Zhang, X., He, M., & Xu, J. (2026). When personalization misleads: Understanding and mitigating hallucinations in personalized LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 8041–8060). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.findings-acl.395>
- Telakivi, P. (2026). Remembering with AI: From distributed memory to AI-curated and human-AI co-memory. *Review of Philosophy and Psychology*. Advance online publication. <https://doi.org/10.1007/s13164-026-00815-1>
- 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>
- Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review*, 2019(2), 494–620. <https://doi.org/10.7916/cblr.v2019i2.3424>
- Warshaw, J., Matthews, T., Whittaker, S., Kau, C., Bengualid, M., & Smith, B. A. (2015). Can an algorithm know the “real you”? Understanding people’s reactions to hyper-personal analytics systems. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 797–806). Association for Computing Machinery. <https://doi.org/10.1145/2702123.2702274>
- Zhang, X., Wang, Y., Chen, R., Wang, Z., Hou, R., & Liu, Z. (2026). Towards proactive personalization through profile customization for individual users in dialogues. In *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 3221–3240). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.findings-acl.159>
- 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>