When Hallucinations Travel: How Evidence-Like Cues Shift Integration, Verification, and Source Memory in Human–AI Interaction

Shiyun Wang[1] **& Xinjie Zhao**[2]
[1]Faculty of Science, University of Copenhagen
[2]Department of Advanced Interdisciplinary Studies, The University of Tokyo

Abstract

AI hallucinations can reshape cognition when false content becomes useful for decisions. We test how AI-generated false claims become usable in reasoning and difficult to separate from their source. Across two preregistered studies (Study 1: $N = 295$; Study 2: $N = 242$), participants made a medication choice after reviewing standardized information and receiving an AI reply varying in hallucination severity. Study 1 shows that stronger hallucinations shifted understanding of a risk relation, changed onset-duration weighting, and increased source confusion, especially for a fabricated causal mechanism. Study 2 traced this pathway in speech: stronger hallucinations increased uptake during decision verbalization, reduced critical evaluation and reliance on original evidence, and weakened source reasoning during interaction recall. Later transmission and post hoc reflection were comparatively bounded. These findings locate hallucination risk at the transition from exposure to reason construction, where AI-originated content becomes useful in participants' explanations while provenance becomes difficult to track.

Keywords: human–AI interaction; hallucination transmission; source monitoring; decision making; misinformation effects

Introduction

People increasingly use conversational AI not only to obtain information, but to build reasons under uncertainty (Bonaccio & Dalal, 2006; Dietvorst et al., 2015; Lee & See, 2004; Logg et al., 2019; Parasuraman & Riley, 1997; Sperber et al., 2010; Yaniv, 2004). False advice is therefore risky not merely because it is false, but because fluency can make a fabricated detail function like evidence. Once such a claim helps explain a tradeoff, it can alter what users weigh and later make the source harder to track (Alter & Oppenheimer, 2009; Ji et al., 2023; Johnson et al., 1993; Loftus, 2005; Maynez et al., 2020; Rozenblit & Keil, 2002). Even relevant prior knowledge can leave people vulnerable to familiar-sounding claims (Fazio et al., 2015). We ask how hallucinated content embedded in fluent conversational advice enters reasoning, changes criteria for justification, and later becomes difficult to separate from its AI source. We link outcome measures to spoken traces of reason construction.

Dual-process accounts distinguish careful evaluation of argument structure from faster reliance on credibility cues such as authority, fluency, or formatting (Chaiken, 1980; Metzger, 2007; Petty & Cacioppo, 1986). In conversational AI settings, this distinction matters because users may incorporate advice into their own reasons rather than simply accept or reject it.

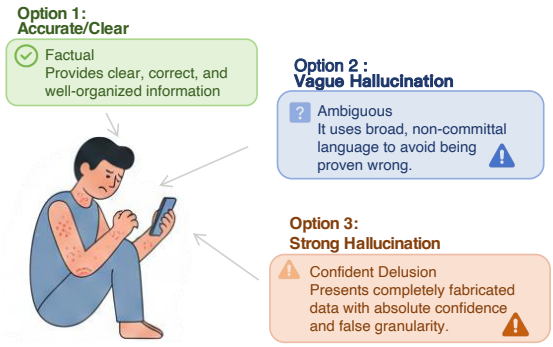

Figure 1: Schematic overview of the three fixed assistant-response conditions, ranging from leaflet-consistent evidence to vague and strong hallucinations.

If integration and verification diverge, AI advice should leave different traces in adoption and source attribution.

Figure 1 summarizes the response manipulation. It builds on work showing that algorithmic advice can be appreciated, avoided, or monitored depending on how system output supports human oversight (Dietvorst et al., 2015; Hancock et al., 2011; Lee & See, 2004; Logg et al., 2019; Parasuraman & Riley, 1997). After observing an error, users can persistently discount an algorithm's recommendations (Dietvorst et al., 2015); designing for appropriate reliance therefore aims to keep users between over-reliance and undue rejection by supporting monitoring and calibration. Conversational AI sharpens this tension because explanation-like replies can carry persuasive false detail before any error is apparent (Hackenburg et al., 2025; Salvi et al., 2025).

Together, these literatures suggest that hallucinations may travel because interaction first helps users build a coherent account and only later invites item-by-item checking. Outside explicit fact-checking, model outputs can serve as explanatory scaffolds while truth evaluation of individual details is delayed (Ji et al., 2023; Steyvers et al., 2025). If so, transmission should appear as relational encoding (Einstein & Hunt, 1980), gist-like memory (Brainerd & Reyna, 2002), source-monitoring difficulty (Johnson et al., 1993), and fast-default versus slow-correction dynamics (Evans & Stanovich, 2013). Postdecision preference construction adds a final implication: recall and explanation are not neutral readouts, but moments when users may consolidate or retell reasons made available

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

during the AI exchange (Enisman et al., 2021; Svenson, 2006).

A transfer-appropriate processing perspective provides a constraint on this prediction: cues that organize reasoning during interaction may later guide retrieval and evaluation, even when item-level provenance is fragile (Craik & Tulving, 1975; Morris et al., 1977; Nairne, 2002). We ask two questions. First, do hallucinated AI claims alter task-specific beliefs, decision criteria, and source memory in a controlled decision setting? Second, can spoken process traces reveal whether those claims enter participants' own reason construction before later source confusion emerges? Across two preregistered studies, we varied hallucination severity in conversational AI responses and tested whether false claims became available as reasons in participants' own reasoning. We then combine outcome measures with verbal reasoning evidence to test a reason-construction account of hallucination transmission. On this account, hallucinations matter most when they are not merely accepted as claims, but converted into usable reasons.

Study 1

Rather than treating hallucinations as isolated false statements, we define *hallucination transmission* as a three-stage cognitive integration: a false claim restructures the user's task model, alters decision-criterion weighting, and loses its source attachment to the AI (Ji et al., 2023; Johnson et al., 1993; Maynez et al., 2020). Fluent rationales accelerate this sequence by making an erroneous premise functionally useful before verification is attempted (Lewandowsky et al., 2012). Study 1 tested this in a controlled medication-choice setting: participants reviewed standardized drug profiles and then received a conversational AI reply at one of three hallucination-severity levels. If transmission occurred, its signature should extend beyond the final choice—altering a task-relevant belief, shifting criterion use, and fragmenting source attribution. We predicted that stronger hallucinations would distort understanding of a key risk attribute, realign attribute weighting, and elevate source confusion for hallucinated claims, with confusion persisting to a delayed post-test (Johnson et al., 1993). Full study protocol was registered at the Center for Spatial Information Science (CSIS), The University of Tokyo.

Participants. We recruited Chinese adult participants via Credamo,[1] an online survey and crowdsourcing platform widely used in China for behavioral research. Of $N = 335$ individuals who entered the study, $N = 295$ remained after excluding incomplete or noncompliant submissions; each participant completed the study once. Participation took approximately 15 minutes, and ethical approval was obtained from The University of Tokyo.

Design, Materials, and Procedure. *Design and manipulations.* A 3×2 between-participants design crossed *hallucination severity in the assistant reply* (three levels) with *decision stance* (two levels). Hallucination severity was varied through three fixed versions of the assistant reply: a fully leaflet-consistent reply (V1), a mildly distorted reply containing overconfident drift and partial fabrication (V2), and an explicitly hallucinated reply that introduced non-verifiable or fabricated claims, including causal assertions about drowsiness risk (V3) (Ji et al., 2023; Maynez et al., 2020). For decision stance, participants either made the choice *for themselves* in an imminent self-medication situation, or composed advice *for a close peer (e.g., a roommate)* facing the same symptoms but in a more severe condition. The peer condition asked participants to act as an "advisor" with greater responsibility for another person's outcome (Tversky & Kahneman, 1981; Yáñez, 2012). Decision stance was included to test whether social responsibility framing would moderate hallucination transmission. We expected any stance-related difference to appear most clearly in responsibility-related or meta-level measures, rather than necessarily disrupting the core content-uptake pathway.

Decision task and stimulus materials. All participants read a short vignette describing an acute, time-pressured symptom-management scenario. They then saw two over-the-counter *medicines* (Pill A vs. Pill B) in a standardized, label-like format. Both medicine profiles required a tradeoff across multiple attributes (e.g., onset speed, duration of relief, and likelihood of drowsiness). Neither option was globally dominant, so participants had to decide which dimension(s) mattered most. After reviewing both profiles, participants made an initial choice and viewed the assistant reply for their assigned hallucination-severity condition. Manipulated claims appeared inside fluent, decision-oriented justification, mirroring how hallucinations can appear as persuasive "reasons" in interactive assistance rather than as isolated false facts (Lee & See, 2004; OpenAI et al., 2024).

Task sequence and recorded outputs. Participants completed the study in a fixed order (Fig. 2 for an overview). They first read the vignette and inspected both medicine profiles, then indicated an initial choice and rated choice certainty. Immediately after reading the assistant reply, participants completed three measures: (i) a comprehension check about which medicine was more likely to cause drowsiness, (ii) a decision-policy item asking which attributes mattered most (e.g., onset, duration, drowsiness risk, side effects, price), and (iii) immediate source judgments (T1) for three target statements, indicating whether each claim came from the medicine label, the assistant, or another source (Lewandowsky et al., 2012). Participants then completed brief post-task items assessing scenario realism and responsibility attribution, and finally repeated the same source-judgment items (T2) to test whether any source confusion persisted after a short within-session delay. This sequence made the pathway observable across belief, criterion, and source measures without relying on a single endpoint.

[1] https://www.credamo.com/

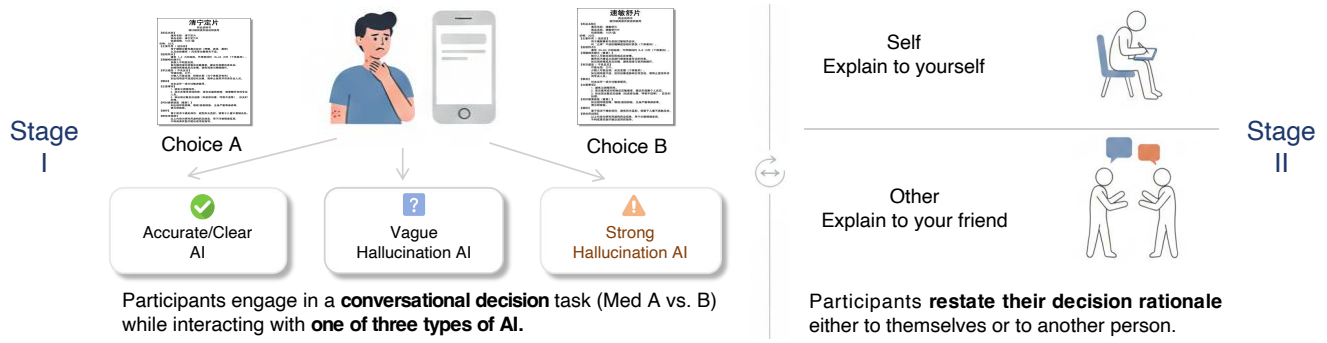

Figure 2: Study-flow schematic. Participants first made a medication choice after reviewing Pill A/Pill B information and one of three assistant-response conditions, then re-articulated their decision rationale either privately or to another person.

Results

Continuous outcomes were analyzed with factorial ANOVA (assistant-response condition × decision stance) and categorical outcomes with Pearson χ^2 tests; effect sizes are reported as η^2 or Cramer's V (Cohen, 1988). Source judgments were coded as AI-source versus non-AI attribution, and each decision attribute as selected versus not selected. Because the outcomes were theoretically linked, we emphasize the predicted cross-measure pattern rather than isolated marginal p-values. Pathway here refers to an ordered evidential profile across belief, criterion, and source measures.

Checks and ancillary outcomes. Participants rated the vignette as realistic (overall realism $M = 5.72$, $SD = 0.95$ on a 1–7 scale), with comparable realism across assistant-response condition, $F(2, 289) = 0.28$, $p = .755$, $\eta^2 = .002$, and decision stance, $F(1, 289) < 0.01$, $p > .999$. Responsibility attribution was comparable across assistant-response conditions, $\chi^2(8) = 13.15$, $p = .107$. Decision stance showed a marginal trend, $\chi^2(4) = 9.01$, $p = .061$, consistent with the possibility that participants deciding for themselves were slightly more inclined to attribute responsibility to themselves. Advice-to-friend choices showed only a marginal condition trend, $\chi^2(6) = 10.70$, $p = .098$, suggesting that this brief behavioral-intention prompt was less sensitive than belief- and source-based measures in the current sample. Together, these checks make a broad demand or engagement account less plausible: the manipulation changed interpretation of the targeted risk and its source more clearly than realism, responsibility allocation, or general advice behavior.

Belief updating: decision certainty and sedation-risk comprehension. Decision certainty varied by assistant-response condition, $F(2, 289) = 7.39$, $p < .001$, $\eta^2 = .048$, but the pattern was modest: V3 produced lower certainty ($M = 5.56$, $SD = 0.97$) than V1 ($M = 5.92$, $SD = 0.88$) and V2 ($M = 5.98$, $SD = 0.82$) on the 1–7 scale, with small decision-stance effects ($p = .278$). Belief change was clearest for the manipulated risk relation. Participants in V1 correctly identified Pill B as more likely to cause drowsiness in 73.0% of cases; this fell to 44.3% in V2 and 19.4% in V3, $\chi^2(2) = 57.44$, $p < .001$, $V = .441$. A binary logistic model with hallucination severity coded ordinally showed the same graded pattern: each step in severity reduced the odds of correct risk identification (OR = 0.35, 95% CI [0.22, 0.55]). Because the item targeted the risk relation distorted by the assistant's rationale, the result is better read as a change in the participant's task model than as a generic comprehension failure (Lewandowsky et al., 2012). Thus, the belief-level effect was asymmetric rather than global: the clearest movement appeared in the risk relation targeted by the fabricated content, not in higher overall certainty, suggesting that hallucinations can alter a local task model without making the whole decision feel more secure.

Criterion reweighting. Assistant-response condition selectively altered which attributes participants treated as important. Selections differed strongly for *fast onset*, $\chi^2(2) = 62.27$, $p < .001$, $V = .459$, and *duration* ("how long the effect lasts"), $\chi^2(2) = 55.54$, $p < .001$, $V = .434$. In V3, participants were much more likely to weight quick onset (77 of 98) and less likely to weight lasting duration (33 of 98) than in V1 (30 of 100 and 74 of 100, respectively). Drowsiness-risk weighting showed only a trend, $\chi^2(2) = 5.09$, $p = .078$, $V = .131$. Side effects, $\chi^2(2) = 0.90$, $p = .638$, and price, $\chi^2(2) = 3.86$, $p = .145$, remained comparatively stable even though these options were available in the same decision-policy item. Reweighting concentrated on attributes licensed by the hallucinated rationale: the strongest hallucination described Pill B as quick and less likely to carry drowsiness into the next morning, so the policy shift appeared most clearly in speed and duration weighting rather than in generic side-effect or price options. Decision-stance effects on attribute weighting were small (all $p \geq .224$), reinforcing the interpretation that reweighting was driven primarily by assistant-response content rather than by the brief self-versus-peer framing.

Source fragility and persistence. Immediate source judgments (T1) showed targeted evidence of source confusion for hallucinated content. For the claim "Pill A does not cause

drowsiness," AI-versus-non-AI source judgments showed a trend by assistant-response condition, $\chi^2(2) = 4.69$, $p = .096$, $V = .126$. For "Pill B may cause drowsiness," condition differences were small, $\chi^2(2) = 0.52$, $p = .771$, $V = .042$. This pattern is informative: the clearest immediate effect appeared for the mechanistic claim "Faster onset implies greater drowsiness," $\chi^2(2) = 7.67$, $p = .022$, $V = .161$, indicating that fabricated causal structure was especially likely to be detached from its AI origin (Johnson et al., 1993; Lewandowsky et al., 2012; Rozenblit & Keil, 2002). Decision-stance effects on T1 source judgments were small for all claims (all $p \geq .264$; all $V \leq .078$), suggesting that immediate confusion was primarily driven by hallucination severity rather than stance. At post-test (T2), the mechanistic claim again showed the clearest source-confusion pattern, $\chi^2(2) = 6.30$, $p = .043$, $V = .146$, compared with smaller condition differences for "Pill A does not cause drowsiness," $\chi^2(2) = 4.06$, $p = .131$, $V = .117$, and "Pill B may cause drowsiness," $\chi^2(2) = 0.45$, $p = .798$, $V = .039$. This persistence is theoretically diagnostic: a causal claim can organize separate details into an explanatory relation, making it easier to retrieve as part of the task model and harder to reattach to the assistant (Johnson et al., 1993; Roediger & McDermott, 1995). Criterion change therefore appeared less as a direct increase in drowsiness-risk weighting than as a reorganization of the speed-duration tradeoff that made the hallucinated recommendation usable. Source fragility was similarly claim-specific: the most persistent confusion concerned the fabricated causal relation, not medication-related content in general.

Study 2

Study 1 demonstrated that hallucinated claims embedded in an AI reply propagate into users' decision-relevant cognition: more severe hallucinations altered what participants believed about key attributes, how they weighted those attributes in their decision policy, and how reliably they tagged the source of contested statements. Study 2 builds on this foundation but targets a critical remaining gap—*how* hallucinated content becomes assimilated into a participant's reasoning stream as the decision unfolds. Rather than relying on end-of-task questionnaires alone, Study 2 introduces a spoken-response, process-tracing approach in which participants verbalize their reasoning at five predefined moments (baseline, decision verbalization, interaction reconstruction, social explanation, post hoc reflection), with the resulting transcripts coded for uptake, evaluative stance, and source-related reasoning. This design allows us to distinguish whether downstream belief change reflects shallow endorsement versus deeper integration into self-generated explanations, and to test whether such integration is accompanied by reduced critical evaluation and weakened provenance tracking across successive stages.

Method

Participants. We recruited 300 adult participants from Credamo; the final participant-level process table comprised $N = 242$ respondents with usable grouped spoken-response data after applying the same recruitment, screening, compensation, and procedural standards as in Study 1.

Design and materials. Study 2 preserved the overarching task structure and core materials from Study 1 (the same vignette, leaflet-like medicine profiles, and AI assistant reply manipulation), but replaced key written responses with *spoken* responses. Participants completed five short oral modules that separated online decision construction from later reconstruction and expression: $A0$ (baseline orientation), $A1$ (online decision verbalization), $A2$ (interaction/content reconstruction), $A3$ (explanation to another person), and $A4$ (post hoc reflection on the decision and the interaction). In this ordering, A2–A4 are postdecision moments in which earlier reasons can be reconstructed, consolidated, or bounded rather than simple repetitions of A1.

Transcription and structured coding. Spoken responses were transcribed and converted into quantitative process indicators via a pre-specified LLM-assisted coding pipeline: module-specific prompts applied JSON-schema-defined construct definitions—developed from human-readable categories and pilot review—to each transcript, extracting planned indicators rather than discovering categories post hoc, and yielding a $1{,}209 \times 48$ typed coding table. Quality checks confirmed nonempty metric coverage across all five modules, with 196–212 usable rows per post-baseline module. Because transcripts could reveal condition-specific content, these ratings are not treated as blind coding; evidential weight rests on planned, converging module-level patterns. Main constructs were defined by module and theoretical role: A1 uptake indexed use of AI-originated false or drift content; A1 critical evaluation indexed questioning, checking, or qualifying the assistant response; A1 leaflet-consistent evidence indexed reliance on the original medicine profiles; A2 source-attribution accuracy and source-confusion evidence indexed provenance tracking; A3 source-disclosure accuracy indexed social provenance transparency; and A4 suspicion/checking items indexed retrospective calibration.

Table 1: A1 decision-construction indicators by hallucination-severity condition (means, *SD* in parentheses; 1–7 scales). V1 = consistent; V2 = mild drift; V3 = hallucination. All $F(2,239)$ tests $p < .001$; stance effects $|d| < .10$.

Measure	V1	V2	V3	F	η^2
Hallucination uptake	1.00 (.00)	3.22 (1.19)	5.47 (1.37)	362.30	.752
Critical evaluation	4.47 (.80)	3.83 (.81)	3.27 (.96)	39.75	.250
Original-evidence reliance	4.56 (1.03)	3.93 (1.34)	3.13 (1.30)	27.02	.184

Results and Discussion

Baseline equivalence. We first verified that random assignment produced comparable baseline profiles. Across eight pre-test indicators (e.g., AI use frequency, baseline trust,

anthropomorphism tendency, verification habits), assistant-response and decision-stance comparisons were small (all $|t| < 1.28$, $p > .20$, $|d| < .17$), supporting the interpretation that downstream effects reflect the experimental interaction rather than pre-existing group differences.

Decision construction (A1): hallucination uptake and evaluative processing. At decision construction, the key question was whether hallucinated content entered participants' own reasoning. The A1 verbalization module yielded the strongest effects of hallucination severity, with decision stance negligible across the core indicators, shown in Table 1. For *hallucination uptake evidence*, the evidence-consistent condition showed a degenerate distribution at the minimum score (all participants received a score of 1, $M = 1.00$, $SD = 0.00$). Uptake increased sharply under mild drift ($M = 3.22, SD = 1.19$) and explicit hallucination ($M = 5.47$, $SD = 1.37$), and assistant-response condition explained the vast majority of variance, $F(2, 239) = 362.30$, $p < .001$, $\eta^2 = .752$. Critical evaluation declined monotonically with hallucination severity (V1: $M = 4.47$, $SD = 0.80$; V2: $M = 3.83, SD = 0.81$; V3: $M = 3.27, SD = 0.96$), with a robust condition effect, $F(2, 239) = 39.75, p < .001, \eta^2 = .250$, consistent with a shift from evaluative to more accepting processing as hallucination severity increased. Reliance on leaflet-consistent evidence showed the same graded pattern (V1: $M = 4.56$, $SD = 1.03$; V2: $M = 3.93$, $SD = 1.34$; V3: $M = 3.13$, $SD = 1.30$), $F(2, 239) = 27.02$, $p < .001$, $\eta^2 = .184$, indicating reduced use of non-AI evidence as hallucination intensified. Reasoning elaboration also differed by assistant-response condition (V1: $M = 5.20$, $SD = 0.82$; V3: $M = 4.63$, $SD = 0.74$), $F(2, 239) = 10.53$, $p < .001$, $\eta^2 = .081$, while AI information adoption remained high but still increased with hallucination severity (V1: $M = 5.46$, $SD = 0.78$; V3: $M = 5.92$, $SD = 1.02$), $F(2, 239) = 6.72$, $p = .001$, $\eta^2 = .053$. A1 was diagnostic as a cluster: uptake rose as verification markers declined (all decision-stance effects $|d| < .10$), localizing reason construction with weakened verification.

Interaction recall (A2): source reasoning, confusion, and meta-level stance effects. At interaction recall, we tested whether uptake left a source-monitoring trace. When participants later reconstructed the interaction, hallucination severity continued to shape source-related and integration indicators. Source attribution accuracy declined with hallucination severity (V1: $M = 4.04$, $SD = 0.95$; V2: $M = 3.43$, $SD = 0.81$; V3: $M = 3.06$, $SD = 0.98$), $F(2, 239) = 23.01$, $p < .001$, $\eta^2 = .161$, indicating poorer tracking of where key claims came from as AI content became more distorted. Recall accuracy for AI content showed the same pattern (V1: $M = 4.29$, $SD = 1.20$; V3: $M = 3.13$, $SD = 1.11$), $F(2, 239) = 23.85$, $p < .001$, $\eta^2 = .166$, suggesting that hallucination severity disrupted reconstruction of what the AI actually said. Source confusion evidence increased from V1 ($M = 4.11$, $SD = 0.78$) to V2 ($M = 4.54$, $SD = 0.61$) to V3 ($M = 4.62$, $SD = 0.88$), $F(2, 239) = 10.27$, $p < .001$, $\eta^2 = .079$. An interaction between assistant-response condition and decision stance, $F(2, 236) = 3.30$, $p = .039$, indicated that stance can modulate how confusion manifests under distorted outputs. Information integration quality decreased as hallucination severity increased (V1: $M = 4.17$, $SD = 0.99$; V3: $M = 3.35$, $SD = 0.93$), $F(2, 239) = 16.22$, $p < .001$, $\eta^2 = .119$. Unlike A1, decision stance produced two reliable meta-level effects in A2. Advising a roommate yielded greater interaction depth ($M = 2.36$, $SD = 1.41$) than deciding for oneself ($M = 1.75$, $SD = 1.23$), $t(240) = -3.55$, $p < .001$, $d = -0.46$. Deciding for oneself showed higher uncertainty acknowledgment ($M = 4.60$, $SD = 1.44$) than advising a roommate ($M = 4.26$, $SD = 1.20$), $t(240) = 1.99$, $p = .048$, $d = 0.26$. A2 located the memory trace of uptake: participants reconstructed the interaction with weaker boundaries among leaflet evidence, assistant additions, and their own inferences, while decision stance operated at the meta-level.

Downstream expression (A3) and post hoc reflection (A4). To test whether private uptake generalized into later expression or self-evaluation, A3 and A4 examined downstream communication and reflection. During downstream expression (A3), assistant-response condition effects were small for the main expression indicators (all $p > .12$; $\eta^2 < .02$), including AI-content transmission, source disclosure accuracy, authority claims, explanatory accuracy, inclusion of warnings, and perceived persuasiveness. A3 suggests that social explanation may flatten severity gradients because AI provenance was often omitted regardless of condition: 57.9% of participants omitted AI as an information source, and only 1.2% provided complete disclosure. Post hoc reflection (A4) added a metacognitive boundary: among seven reflection outcomes, decision-process evaluation differed by assistant-response condition (V1: $M = 5.58$, $SD = 0.65$; V3: $M = 5.28$, $SD = 0.58$), $F(2, 239) = 4.57$, $p = .011$, $\eta^2 = .037$, whereas post-decision confidence, perceived AI reliability, information sufficiency, post hoc suspicion, future AI-use intention, and checking intention were comparatively stable. This late-stage profile sharpens the theoretical claim: hallucinated content was available during private decision construction, while public retransmission and retrospective self-evaluation showed a more bounded pattern—consistent with postdecision consolidation processes that stabilize coherence while item-level provenance recedes (Enisman et al., 2021; Svenson, 2006). Study 2 thereby localized the pathway temporally: most potent during reason construction and interaction reconstruction, attenuated during later transmission and self-evaluation.

General Discussion

Across both studies, hallucination effects were selective rather than globally persuasive: severity shifted local task beliefs, the

speed-duration tradeoff, and provenance judgments, and appeared in spoken reasoning when participants constructed or reconstructed explanations, while public retransmission and broad self-evaluation showed a bounded profile. The empirical contribution is therefore not that hallucinations increase trust or persuasion in general, but that they become consequential when recruited as reasons.

We interpret this selectivity as *reason assimilation*: false content becomes consequential when it organizes a choice, explains a tradeoff, or justifies a preference. Once used in that way, the claim is stored less as an utterance from the assistant than as part of the user's task model, producing the observed content-source dissociation and converging with gist/verbatim distinctions, source-monitoring error, and postdecision preference construction (Brainerd & Reyna, 2002; Johnson et al., 1993; Svenson, 2006).

Positioning this pattern against broader misinformation research: once an explanation is incorporated into a mental model, it tends to guide judgment even when its evidential basis is later questioned—what Lewandowsky et al. (2012) call the *continued-influence effect*. What is distinctive in conversational AI settings is the interactional channel. Dialogue format encourages coherence-seeking and pragmatic acceptance, increasing the likelihood that a fluent rationale is adopted as usable reasoning support rather than quarantined as an externally checkable claim—and this process operates most powerfully before verification is ever attempted. The present findings specify where uptake enters the decision lifecycle and link source-monitoring fragility to that prior constructive moment.

Most diagnostic was the fabricated causal mechanism, which supplied relational structure rather than isolated content. By linking onset speed to drowsiness risk, it reorganized the decision space, making it easier to recruit during reasoning and harder to reattach to its AI source—consistent with the illusion of explanatory depth as a provenance-erosion mechanism (Rozenblit & Keil, 2002).

Study 1 and Study 2 provide complementary forms of evidence. Study 1 shows what changed after exposure; Study 2 shows how the content appeared in spoken rationales while critical evaluation and reliance on the original evidence declined. Their convergence supports a content-source dissociation in which AI-originated material becomes available as a reason while its provenance becomes unstable. The dual-channel framing—integration versus verification—functions as a theoretical organizing construct rather than a directly tested mechanism; a stronger causal decomposition would require manipulations that selectively disrupt one channel. Content-specificity in the reweighting results—concentrated on speed and duration rather than peripherally distributed—distinguishes selective transmission from generic persuasion. The null effects of decision stance on core uptake and source measures indicate that brief social-responsibility framing does not buffer reason assimilation, suggesting that the uptake process operates prior to the level at which social accountability typically intervenes.

For interface design, the relevant intervention point is reason construction. Systems should preserve the relation among claims, sources, and user reasoning operations: what came from the original evidence, what the assistant inferred or added, and what remains unsupported. Provenance-sensitive prompting, structured evidence comparison, and disconfirming checks embedded in the interaction may be more useful than generic warnings delivered after uptake.

Several limits follow from the design. The medication vignette enabled controlled manipulation but does not estimate field prevalence or magnitude in open-ended AI use; the results identify a transmission mechanism rather than establish base rates. Fabrications were relatively salient, and whether the same process operates for subtler hallucinations—plausible numerical errors or unsourced causal claims—remains an open question. Fixed replies increased internal validity while packaging severity with confidence, coherence, and recommendation framing; the active ingredient of transmission cannot be fully decomposed without factorial variation of these features. Multiple converging outcomes were tested without a family-wise correction framework, and both samples were drawn from Chinese adults via Credamo; cross-cultural generalization requires caution given documented variation in AI trust calibration and collectivist decision norms (Hancock et al., 2011; Lee & See, 2004). Study 2 also reveals a dissociation between private uptake and public expression: participants reliably incorporated hallucinated content during reasoning, yet did not show elevated retransmission in the social explanation module—distinguishing reason construction from downstream communication and marking social disclosure as a distinct target for future work. Study 2 used schema-guided LLM-assisted coding; future work should pair this pipeline with per-measure human reliability calibration, test longer retention intervals and subtler manipulations, and examine moderators such as prior knowledge, AI literacy, and domain familiarity.

Acknowledgments

This work was supported by JST BOOST through the University of Tokyo BOOST NAIS program, Japan Grant Number JPMJBS2418. This research was also partially supported by the Recommendation Program for Young Researchers and Woman Researchers, Information Technology Center, The University of Tokyo. Part of this work was conducted using mdx resources allocated to project 2025-A-029, "Large Language Model-driven Interpretable Multi-Modal Decision Support System for Disaster Prevention and Mitigation."

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the Tribes of Fluency to Form a Metacognitive Nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Brainerd, C., & Reyna, V. (2002). Fuzzy-Trace Theory and False Memory. *Current Directions in Psychological Science*, *11*(5), 164–169. https://doi.org/10.1111/1467-8721.00192

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*(5), 752–766. https://doi.org/10.1037/0022-3514.39.5.752

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates. https://books.google.co.jp/books?id=gA04ngAACAAJ

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268

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

Einstein, G. O., & Hunt, R. R. (1980). Levels of processing and organization: Additive effects of individual-item and relational processing. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 588–598. https://doi.org/10.1037/0278-7393.6.5.588

Enisman, M., Shpitzer, H., & Kleiman, T. (2021). Choice changes preferences, not merely reflects them: A meta-analysis of the artifact-free free-choice paradigm. *Journal of Personality and Social Psychology*, *120*(1), 16–29. https://doi.org/10.1037/pspa0000263

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

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

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational artificial intelligence. *Science*, *390*(6777), eaea3884. https://doi.org/10.1126/science.aea3884

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3–28. https://doi.org/10.1037/0033-2909.114.1.3

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, *12*(4), 361–366. https://doi.org/10.1101/lm.94705

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

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020, July). On Faithfulness and Factuality in Abstractive Summarization. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.173

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, *58*(13), 2078–2091. https://doi.org/10.1002/asi.20672

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Nairne, J. S. (2002). The myth of the encoding-retrieval match. *Memory*, *10*(5–6), 389–395. https://doi.org/10.1080/09658210244000216

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Zoph, B. (2024, March 4). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs]. https://doi.org/10.48550/arXiv.2303.08774

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of Persuasion. In L. Berkowitz (Ed.), *Ad-*

vances in Experimental Social Psychology (pp. 123–205, Vol. 19). Academic Press. https://doi.org/10.1016/S0065-2601(08)60214-2

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, *9*(8), 1645–1653. https://doi.org/10.1038/s41562-025-02194-6

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic Vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, *7*(2), 221–231. https://doi.org/10.1038/s42256-024-00976-7

Svenson, O. (2006). Pre- and Post-Decision Construction of Preferences: Differentiation and Consolidation. In P. Slovic & S. Lichtenstein (Eds.), *The Construction of Preference* (pp. 356–371). Cambridge University Press. https://doi.org/10.1017/CBO9780511618031.020

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, *211*(4481), 453–458. https://doi.org/10.1126/science.7455683

Yáñez, C. S. (2012). Mercier and Sperber's Argumentative Theory of Reasoning: From Psychology of Reasoning to Argumentation Studies. *Informal Logic*, *32*(1), 132–159. https://doi.org/10.22329/il.v32i1.3536

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13. https://doi.org/10.1016/j.obhdp.2003.08.002