Do Humans and LLMs Diverge in Belief Revision? Evidence from a Bayesian Analysis

Stylianos Loukas Vasileiou[1]*, **Antonio Rago**[2], **Maria Vanina Martinez**[3], **William Yeoh**[4]
New Mexico State University[1], King's College London[2], IIIA-CSIC[3],
Washington University in St. Louis[4]

Abstract

Large language models are increasingly used as proxies for human cognition, but do they reason like humans when beliefs conflict with new evidence? We study this question through *belief revision*, that is, how agents update their beliefs in light of contradictions. Cognitive psychology has shown that humans favor *explanation-based revision*: they first generate explanations for the conflict, such as identifying conditions under which a general rule admits exceptions, and revise accordingly. In this paper, we develop a Bayesian model that decomposes this process into two orthogonal components: *what* to revise and *how broadly* to generalize to similar instances. We first run three human-subject experiments on simple, everyday abstract and grounded scenarios to evaluate how people revise their beliefs, and use this data to test our model's predictions. We then evaluate three frontier LLMs on the same scenarios and show that while LLMs match human rule-targeting rates, they consistently fail to generalize: both humans and LLMs target the directly contradicted beliefs, but humans propagate revisions to related beliefs while LLMs revise only the directly contradicted instance—a gap present in both abstract and grounded settings.

1 Introduction

A growing body of work treats large language models (LLMs) as "silicon subjects", proxies for human cognition that can simulate participant responses in psychological experiments at scale (Aher et al., 2023; Shanahan et al., 2023). If this assumption holds, LLMs could dramatically accelerate cognitive science by replacing costly human-subject studies with scalable model queries. But how far does the analogy extend? Do LLMs produce statistically plausible responses, or do they engage in the same reasoning processes as humans?

We investigate this question through the lens of *belief revision*, that is, how an agent updates its beliefs when new information contradicts them. Belief revision is a foundational problem at the intersection of AI and cognitive science, and it offers a particularly clean test case because there exist competing theoretical predictions about how a rational agent should behave. The dominant principle in AI, dating back to James (1907), is *minimalism*: when faced with a contradiction, make the smallest possible change to restore consistency (Makinson, 1997; Rott, 2000; Fermé et al., 2024). Yet a sustained line of work in cognitive psychology has argued that humans do not follow this principle, but, instead, they seek *explanatory understanding*, generating causal explanations for the conflict that lead to broader, seemingly non-minimal revisions (Elio & Pelletier, 1997; Walsh & Johnson-Laird, 2009; Khemlani & Johnson-Laird, 2013).

In this paper, we study belief revision from a computational standpoint (Marr, 1982) by developing a Bayesian model that makes quantitative predictions. Our model decomposes belief revision into two orthogonal components: (1) *what* to revise—whether the agent targets a *conditional* statement, such as a general rule, or a *categorical* statement, such as an assertion about a particular instance, governed by prior beliefs about rule defeasibility and

*Correspondence: stelios@nmsu.edu

categorical reliability; and (2) *how broadly* to generalize the revision across related beliefs, governed by a generalization parameter. We test these predictions in *abstract* scenarios involving general rules (e.g., "if people are worried, they find it hard to concentrate"), and *grounded* scenarios involving instantiated rules as specific claims about particular individuals and contexts (e.g., "if presentation worry occurs, then Alice loses concentration"). We then evaluate both humans and state-of-the-art LLMs against these predictions.

Our central finding is a disagreement between these two components. Humans and LLMs *agree* on what to revise, that is, faced with a contradiction, both overwhelmingly target rules rather than the categoricals. However, they *disagree* on how broadly that revision propagates. Human revision propagates across structurally related beliefs, whereas LLM revision stays confined to the directly contradicted instance, in both scenarios. LLMs therefore reproduce the targeting pattern without the propagation that generates it in humans, which is a *functional* rather than a *literal* gap (Riemer et al., 2025; Mahowald et al., 2024), in which the surface behavior is matched while the process that produces it is not.

Our main contributions are: (1) **a Bayesian model of belief revision** (Section 3) that yields a single closed-form equation (Equation 4) parameterized by the number of contradicted rules and categoricals, together with a generalization parameter γ; (2) **a human baseline from three experiments** (Section 4) that calibrates the model's priors and establishes the human generalization rates; and (3) **a human–LLM divergence in belief revision** (Section 5) showing that LLMs match human rule-targeting rates but fail to propagate revisions to related beliefs.

2 Related Work

Belief revision in cognitive science. The principle of minimalism—that a rational agent should make the smallest change to restore consistency—has deep roots in philosophy (James, 1907) and formal AI, most notably the AGM framework (Alchourrón et al., 1985; Fermé et al., 2024; Rott, 2000). However, cognitive psychologists have consistently challenged this as a descriptive account of human reasoning. Elio & Pelletier (1997) showed that people prefer to modify general conditionals rather than retract specific categoricals, a finding replicated by Politzer & Carles (2001) and extended by Byrne & Walsh (2002). Walsh & Johnson-Laird (2009) and Khemlani & Johnson-Laird (2013) demonstrated that people generate "disabling conditions" (explanations that make a rule defeasible) rather than making minimal, arbitrary changes. Our work provides the first formal Bayesian model that explains *why* rule revision is favored and, crucially, directly contrasts human behavior with that of frontier LLMs.

LLMs as cognitive models. Recent work has explored LLMs as models of human cognition, from evaluating them on cognitive tasks (Binz & Schulz, 2023) to finetuning on behavioral data (Binz & Schulz, 2024) and training on computationally equivalent tasks (Zhu et al., 2025), with increasing success at predicting human behavior (Binz et al., 2025). However, LLMs show systematic deviations: they can replicate aggregate patterns but often fail on individual variation (Aher et al., 2023). A key distinction is between *literal* and *functional* competence (Riemer et al., 2025; Kosinski, 2024), where an LLM may predict behavior without performing the underlying reasoning. Our work contributes a new evaluation domain, that of belief revision, where this contrast is surfaced.

Belief revision and LLMs. Two recent papers address belief revision in LLMs directly. Hase et al. (2024) critique the model editing literature through the lens of belief revision theory, arguing that LLMs may lack the coherent, revisable belief systems assumed by classical frameworks like AGM, and identifying fundamental challenges for specifying how rational belief revision should work in language models. Wilie et al. (2024) introduce the Belief-R benchmark to evaluate whether LLMs can update their reasoning when presented with information that defeats a prior modus ponens conclusion, finding that most models struggle with this form of nonmonotonic revision. Our work differs from both, insofar as

we test whether they can simulate *human* belief revision patterns using real human data—a task relevant to cognitive modeling and human-AI interaction.

Bayesian models of cognition. Our approach follows the rational analysis tradition (Marr, 1982; Anderson, 1990; Griffiths et al., 2010; 2024), which derives optimal solutions to computational problems under environmental constraints. As Griffiths et al. (2023) argued, Bayesian models complement neural network approaches by providing interpretable accounts of *why* agents behave as they do. Our model exemplifies this, using just three parameters to yield testable predictions that are directly interpretable in terms of the cognitive priors that drive behavior.

3 A Bayesian Model of Belief Revision

We develop a Bayesian model that predicts *what* an agent revises and *how broadly* they generalize that revision in the following context. Consider the following scenario, where an agent holds two beliefs:

R_1: *If people are worried, then they find it difficult to concentrate.* (conditional)
F_1: *Alice was worried.* (categorical)

Together, R_1 and F_1 predict that Alice found it difficult to concentrate. But the agent now learns that φ: *Alice did **not** find it difficult to concentrate*. The agent has two options, either reject the conditional ("worry does not always impair concentration—perhaps Alice has effective coping strategies"), or reject the categorical ("Alice was not actually worried"). The first option targets the conditional statement, while the second targets the categorical statement.

More generally, we assume an agent maintains a knowledge base $\mathcal{K} = (\mathcal{R}, \mathcal{F})$, where $\mathcal{R} = \{R_1, \ldots, R_n\}$ is a set of *conditional statements* (generalizations) of the form $C_i \rightarrow E_i$ ("if C_i then E_i") and $\mathcal{F} = \{F_1, \ldots, F_m\}$ is a set of *categorical statements* (assertions about specific instances). When R_i and F_j jointly predict $E_i(a)$ but the agent observes an *epistemic input* $\varphi = \neg E_i(a)$, the knowledge base is inconsistent and must be revised. We denote by s the number of rules whose predictions are contradicted by φ, and by q the number of distinct categoricals involved in the contradiction (i.e., categoricals F_j that, together with some contradicted rule, jointly predict an outcome negated by φ). In all experiments reported in this paper, $q = 1$; we present the model for general q for completeness. For brevity, we refer to conditional statements as *rules* and to categorical statements as *categoricals* hereafter.

We model this revision decision as Bayesian inference, where the agent maintains uncertainty about whether each of its beliefs is reliable, and updates upon encountering a contradiction. Specifically, we associate two latent properties with the agent's beliefs:

- **Rule universality (θ_r).** Each rule R_i is either *strictly universal*, i.e., it holds without exception, or *defeasible*, i.e., it admits exceptions. The prior probability that a rule is universal is $\theta_r \in (0, 1)$. A high θ_r (e.g., 0.9) means the agent believes rules are almost always exceptionless; a low θ_r (e.g., 0.2) means exceptions are readily expected.
- **Categorical reliability (θ_f).** Each categorical F_j is either *correct* or *incorrect*. The prior probability that a categorical is correct is $\theta_f \in (0, 1)$. A high θ_f (e.g., 0.8) means the agent trusts that asserted categoricals typically hold.

Let $U_i \in \{0, 1\}$ indicate whether rule R_i is universal ($U_i = 1$) or defeasible ($U_i = 0$), and let $V_j \in \{0, 1\}$ indicate whether categorical F_j is correct ($V_j = 1$) or not ($V_j = 0$). We assume all U_i and V_j are independent a priori. After observing φ, the agent conditions on φ, eliminating all joint states that are inconsistent with the observation.

Deriving the revision probability. We derive the model's key prediction for the case $s = 1$ (one contradicted rule) and then generalize. In the Alice example, the agent considers three possible states of the world:

1. *Rule defeasible, categorical correct* ($U_1 = 0$, $V_1 = 1$): The rule admits exceptions; Alice was worried but her coping strategies prevented concentration loss. Prior mass: $\theta_f(1-\theta_r)$.
2. *Rule defeasible, categorical incorrect* ($U_1 = 0$, $V_1 = 0$): The rule admits exceptions *and* Alice was not actually worried. Prior mass: $(1-\theta_f)(1-\theta_r)$.
3. *Rule universal, categorical incorrect* ($U_1 = 1$, $V_1 = 0$): The rule holds without exception, so the categorical must be wrong, e.g., Alice was not worried. Prior mass: $(1-\theta_f)\theta_r$.

The fourth combination $U_1 = 1$, $V_1 = 1$ (rule universal, categorical correct) is *eliminated* because it would predict that Alice found it difficult to concentrate, contradicting φ. After conditioning, the posterior is obtained by renormalizing the surviving states. The normalizing constant is $Z = \theta_f(1-\theta_r) + (1-\theta_f)(1-\theta_r) + (1-\theta_f)\theta_r = 1 - \theta_f\theta_r$. The posterior yields three revision outcomes:

$$P(\text{rule only} \mid \varphi) = \frac{\theta_f(1-\theta_r)}{Z} \tag{1}$$

$$P(\text{rule \& cat.} \mid \varphi) = \frac{(1-\theta_f)(1-\theta_r)}{Z} \tag{2}$$

$$P(\text{cat. only} \mid \varphi) = \frac{(1-\theta_f)\,\theta_r}{Z} \tag{3}$$

For the general case with s contradicted rules and q involved categoricals, we assume a *fully crossed* structure: every contradicted rule paired with every involved categorical produces a contradicted prediction. A joint state $(\mathbf{U}, \mathbf{V})$ survives conditioning on φ if and only if, for every such pair, either the rule is defeasible or the categorical is incorrect. The surviving states partition into two cases: (i) all s rules are defeasible, in which case any categorical configuration is consistent (mass $(1-\theta_r)^s$); (ii) at least one rule is universal, which forces all q involved categoricals to be incorrect (mass $[1-(1-\theta_r)^s](1-\theta_f)^q$). The normalizing constant is therefore $Z = (1-\theta_r)^s + [1-(1-\theta_r)^s](1-\theta_f)^q$, and the probability that the agent targets at least one rule is:

$$\boxed{P(\text{rule} \mid \varphi) = 1 - \frac{\theta_r^s\,(1-\theta_f)^q}{(1-\theta_r)^s + [1-(1-\theta_r)^s](1-\theta_f)^q}} \tag{4}$$

where s is the number of contradicted rules and q the number of involved categoricals. When $q = 1$, this reduces to $1 - (1-\theta_f)\theta_r^s / [(1-\theta_f) + \theta_f(1-\theta_r)^s]$, which is the form used throughout this paper. The three-way posterior also generalizes: $P(\text{rule only} \mid \varphi) = (1-\theta_r)^s\,\theta_f^q / Z$, $P(\text{cat. only} \mid \varphi) = \theta_r^s\,(1-\theta_f)^q / Z$, and $P(\text{both} \mid \varphi) = 1 - P(\text{rule only}) - P(\text{cat. only})$.

What the equation predicts. To build intuition, consider the case $s = 1$, $q = 1$ (as in all our experiments). If the agent trusts categoricals highly ($\theta_f = 0.8$) but views rules as moderately defeasible ($\theta_r = 0.5$), then $P(\text{rule} \mid \varphi) = 1 - \frac{0.2 \times 0.5}{0.2 + 0.8 \times 0.5} = 1 - \frac{0.10}{0.60} = 0.83$. When $\theta_f > \theta_r$, agents who trust their categoricals more than the universality of their rules will prefer rule-targeting revision. Moreover, $P(\text{rule} \mid \varphi)$ is monotonically increasing in s when $\theta_f\theta_r < 0.5$ (Appendix A.2): the more rules are contradicted, the more likely the agent is to target them.

Equation 4 and the posterior predict *what* the agent revises. A separate, descriptive parameter characterizes *how broadly* they generalize across related beliefs. When a scenario contains multiple structurally related rules—e.g., rules sharing an antecedent, or ground instantiations of the same abstract rule—a contradiction with one may or may not propagate to the others. We introduce a generalization parameter $\gamma \in [0, 1]$: the probability that each non-contradicted but related rule is also revised. Let N denote the total number of related rules in the scenario. The expected number of rules changed among rule-targeters is:[1]

$$\mathbb{E}[\text{rules changed} \mid \text{rule-targeting}] = s + \gamma(N - s) \tag{5}$$

[1] In practice, we estimate γ from data by rearranging: $\hat{\gamma} = (\bar{n} - s)/(N - s)$, where $\bar{n}$ is the observed mean number of rules revised.

When $\gamma = 0$, rule-targeters revise only the directly contradicted rule(s) (local repair). When $\gamma = 1$, they revise all related rules (full theory-level revision).

Putting it all together, the model yields two orthogonal, testable hypotheses:

> **H1** (Rule-targeting preference): Agents prefer to revise rules over categoricals, i.e., $P(\text{rule} \mid \varphi) > 0.5$.
>
> **H2** (Generalization): Agents that target rules generalize beyond the directly contradicted instance, i.e., $\gamma > 0$.

H1 can be tested via a uniform classification: did the revision target at least one rule? H2 is testable whenever the scenario contains non-contradicted rules related to the contradicted one—either abstract rules sharing an antecedent or multiple groundings of the same abstract rule. Note that part of the rule/categorical asymmetry (H1) is constitutive, i.e., admitting exceptions is part of what makes a rule a rule, and being asserted of a particular instance is part of what makes a categorical a categorical. H1's role, therefore, is to verify that agents engage the task as intended and to calibrate θ_f and θ_r, so that γ can be estimated on a common footing for humans and models.

4 Human Experiments

We conducted three experiments to test H1 and H2. All studies were run on Prolific (Palan & Schitter, 2018) with English-fluent adult participants. Participants were told there were no right or wrong answers.

4.1 Experimental Design

To test the model at different levels of complexity, we designed three types of inconsistency scenarios that vary the number of rules and contradictions.

Type I ($s = 1$). The agent holds one rule R_1 ("if people are worried, they find it difficult to concentrate") and one categorical F_1 (e.g., "Alice was worried"). Together these predict that Alice found it difficult to concentrate. The epistemic input φ contradicts this prediction: "Alice did *not* find it difficult to concentrate." The agent must decide whether to revise R_1, F_1, or both.

Type II ($s = 1$). A second rule R_2 sharing the same antecedent is added, e.g.,"if people are worried, then they have insomnia." The epistemic input still contradicts only R_1's prediction. However, the agent must now additionally decide whether R_2 should also be revised, that is, whether the disabling condition that undermines R_1 also applies to R_2.

Type III ($s = 2$). Here, both rules' predictions are contradicted simultaneously with the epistemic input φ:"Alice did not find it difficult to concentrate *and* did not have insomnia."

Each type includes three distinct scenarios drawn from everyday domains—psychology, physics, and economics—for nine scenarios total per experiment (see Appendix for the full set). Note that in all three experiments, the rules and categoricals were presented as claims attributed to speakers, and only the epistemic input φ was presented as known to be true (Appendix A.1 gives the verbatim instructions and stimuli). Categoricals were therefore open to doubt by construction, as nothing in the framing marked them as incorrigible, and participants were free to reject them.

Classification. Across all three experiments, we apply a uniform *type-based* classification. A response is *rule-targeting* if the participant's revision targets at least one rule (by modifying it (e.g., adding exceptions) or discarding it). It is *categorical-only* if only categoricals are affected. Responses that neither affirm nor deny any belief, or are too vague to classify, are excluded. This classification directly maps onto the model's $P(\text{rule} \mid \varphi)$ from Equation 4.

Statistical methodology. For each participant, we compute the proportion of rule-targeting revisions across their valid responses. We test H1 using a one-sided Wilcoxon signed-rank test on these proportions against a null median of 0.5, with $\alpha = 0.01$. Effect sizes are reported as Cohen's d_z.

4.2 Experiment 1: Self-Generated Explanations

The goal of Experiment 1 is to test H1 in a setting where participants are free to revise their beliefs in any way they see fit, without being given an explanation or a structured response format. This tests whether the preference for rule-targeting revision arises naturally.

Participants & Procedure. We recruited 62 participants (age: 20–64, $M = 37.1$; 30 female, 31 male, 1 other). Participants were presented with each of the nine scenarios. For each one, they were asked to explain in their own words *why* the epistemic input could be true despite the conflicting initial statements.

Coding. Free-text explanations were classified based on whether the explanation targeted a rule (e.g., by invoking a disabling condition, e.g., "Alice has effective coping strategies, so worry did not impair her concentration") or a categorical (e.g., "Alice was not actually worried"). This scheme classified 89% of responses (the remaining 11% were ambiguous and excluded).

Results. Rule-targeting rates were high across all types: 82.0% for Type I, 87.0% for Type II, and 81.4% for Type III (Table 1). All were significantly above chance (all $p < 10^{-7}$, all $d_z > 0.87$), confirming H1. Participants generated disabling conditions (e.g., "Alice may have developed coping mechanisms") rather than questioning the categoricals. Because Experiment 1 uses free-text responses, we can classify what participants target but not how many related beliefs they would revise; accordingly, γ cannot be estimated for this experiment.

4.3 Experiment 2: Provided Explanations

In this experiment, we explore if the preference for rule-targeting revisions persist when the explanation is *provided* rather than self-generated. would revise each statement.

Participants & Procedure. 61 new participants (age: 21–67, $M = 34.8$; 34 female, 27 male). Participants were presented with the same nine scenarios, this time accompanied by an explanation—a disabling condition drawn from the most common rule-targeting explanations produced in Experiment 1 (e.g., "if people have effective coping strategies, then they may still be able to concentrate despite being worried"). For each statement, participants indicated whether they would *keep*, *discard*, or *alter* it. If they chose to alter, they provided a revised version.

Results. Rule-targeting rates remained high: 79.4% (Type I), 94.5% (Type II), and 85.4% (Type III), all significantly above chance (all $p < 10^{-8}$, all $d_z > 0.95$). The similarity to Experiment 1 indicates that the preference does not depend on whether the explanation is self-generated or externally provided. The mean number of belief changes increased with scenario complexity (Table 1).

Because Type II scenarios contain a non-contradicted rule (R_2) sharing the same antecedent as the contradicted rule (R_1), we can test H2 even in abstract settings: $N = 2$, $s = 1$. Among the 155 rule-targeting responses, the estimated generalization parameter is $\hat{\gamma} = 0.42$ (bounds $[0.36, 0.50]$), meaning roughly 42% also revised the non-contradicted rule R_2. Moreover, 12% of rule-targeters also revised the categorical, compared to the model's prediction of $1 - \theta_f = 19\%$ (Appendix A.2). These results show that generalization to related beliefs occurs even with abstract rules, not only in grounded scenarios.

		Classification			H1 ($P > 0.5$)		H2 ($\gamma > 0$)	
Exp.	**Type (s)**	Rule	Cat.	Valid	p	d_z	$\bar{n}$	$\hat{\gamma}$
1	I (1)	132	29	161	1.1×10^{-7}	0.87	—	—
	II (1)	140	21	161	8.6×10^{-11}	1.45	—	—
	III (2)	144	33	177	1.3×10^{-7}	0.95	—	—
2	I (1)	131	34	165	3.9×10^{-8}	0.95	1.18	—
	II (1)	155	9	164	1.7×10^{-12}	3.06	1.63	0.42
	III (2)	129	22	151	1.1×10^{-8}	1.14	1.96	—
3	I (1)	150	7	157	1.1×10^{-12}	2.51	2.34	0.45
	II (1)	150	3	153	6.7×10^{-13}	3.22	3.28	0.33
	III (2)	160	2	162	1.6×10^{-13}	3.25	4.07	0.35

Table 1: Results across all three experiments. $\bar{n}$: mean belief changes per response for Exp. 2 (total changes across all statements); mean rules changed among rule-targeters for Exp. 3. $\hat{\gamma}$: generalization parameter estimated from rule changes (Exp. 2 Type II: $N=2$, $s=1$; Exp. 3: see text for N, s per type).

4.4 Experiment 3: Grounded Scenarios

Experiments 1 and 2 used abstract, general rules (e.g., "if people are worried…"). Experiment 3 tests whether the same patterns hold when rules are *grounded*—instantiated as specific claims about particular individuals and contexts. While Experiment 2 already confirmed H2 for abstract Type II, grounding creates many more instances ($k = 4$ per abstract rule), enabling a finer-grained test of generalization breadth.

Participants & Procedure. 60 new participants (age: 21–74, $M = 34.2$; 33 female, 27 male). Each abstract rule from the previous experiments was instantiated into $k = 4$ specific ground rules by varying the individuals and contexts. Type I scenarios thus contained 4 ground rules and 2 categoricals (6 statements total); Types II and III contained 8 ground rules and 2 categoricals (10 statements). Participants indicated keep/discard/alter for each statement and were also given an explanation (as in Experiment 2).

The epistemic input contradicted one specific ground rule (e.g., "Alice did *not* lose concentration during her presentation"). The key question is whether participants revised only the directly contradicted ground rule, or also generalized to the other instances.

Results. Rule-targeting rates were *higher* than in the abstract scenarios: 95.5% (Type I), 98.0% (Type II), and 98.8% (Type III), all with $p < 10^{-12}$ and $d_z > 2.5$ (Table 1). This finding runs counter to what one might expect from prior count-based analyses, which reported lower "non-minimality" in grounded scenarios. The apparent discrepancy arises because count-based classification conflates rule-targeting (whether a rule was revised at all) with generalization breadth (how many groundings were revised). When these two dimensions are separated, the data reveal that participants are *more*, not less, inclined to target rules when those rules are specific and concrete.

Moreover, among rule-targeters, the mean number of ground rules revised exceeded the number directly contradicted in all types (all $p < 10^{-8}$), confirming $\gamma > 0$. From Equation 5, the estimated generalization parameter is $\gamma = 0.45$ for Type I ($s = 1$, $N = 4$; mean rules changed $= 2.34$), $\gamma = 0.33$ for Type II ($s=1$, $N=8$; mean $= 3.28$), and $\gamma = 0.35$ for Type III ($s=2$, $N=8$; mean $= 4.07$).

The distribution of rules changed reveals a striking pattern. For Type I, rather than a unimodal distribution centered at the mean, we observe a trimodal pattern: 34% of rule-targeters changed only the directly contradicted rule ($\gamma_i = 0$), 30% changed exactly two, and 32% changed all four ($\gamma_i = 1$). This suggests participants follow one of three distinct strategies—local repair, partial generalization, or full theory-level revision—rather than "slightly generalizing" on average (see Appendix A.3 for the full distribution).

Setting	Type (s)	Obs.	Pred.	Δ	Role
Abstract	I ($s=1$)	80.7%	80.7%	0.0	fit
	II ($s=1$)	90.7%	80.7%	10.0	pred.
	III ($s=2$)	83.4%	83.4%	0.0	fit
Grounded	I ($s=1, k=4$)	95.5%	95.5%	0.0	fit
	II ($s=1, k=4$)	98.0%	95.5%	2.5	pred.
	III ($s=2, k=4$)	98.8%	99.0%	0.2	pred.

Table 2: Model predictions versus observed rule-targeting rates. Abstract rates are averaged across Experiments 1 and 2. "Fit" rows were used for parameter estimation; "pred." rows are out-of-sample predictions. Type II has $s = 1$ like Type I, so the model predicts the same rate; the observed gap suggests correlated priors for rules sharing an antecedent (see Limitations).

Therefore, H1 is confirmed across all three experiments and all scenario types: people overwhelmingly prefer to revise rules rather than categoricals, whether working with abstract generalizations (79–95%) or concrete instances (95–99%).[2] H2 is confirmed in both abstract (Experiment 2 Type II: $\gamma = 0.42$) and grounded settings (Experiment 3: $\gamma = 0.33$–0.45): rule-targeters consistently generalize to related beliefs beyond the directly contradicted instance.

4.5 Fitting the Bayesian Model to Human Data

We now estimate the model's parameters and test its predictive accuracy. The procedure has two steps, each using a different subset of the data, so that the model's predictions for the grounded Type III condition serve as a genuine out-of-sample test.

Step 1: Estimating θ_f and θ_r^{abs} from Experiments 1 and 2. We estimate the categorical reliability prior θ_f and the abstract rule universality prior θ_r^{abs} by minimizing the squared error between the model's predictions (Equation 4) and the observed rule-targeting rates for Type I ($s = 1$) and Type III ($s = 2$) in the abstract scenarios, averaged across Experiments 1 and 2. This yields $\theta_f = 0.81$ and $\theta_r^{\text{abs}} = 0.55$. The estimates indicate that people enter these tasks with a strong prior that stated categoricals are correct (81%), but a roughly even prior on whether a general rule admits exceptions (55% universal).

Step 2: Estimating θ_r^{gnd} from Experiment 3, Type I only. We fix $\theta_f = 0.81$ from Step 1—assuming that categorical reliability does not change when rules are grounded—and estimate a separate rule universality prior θ_r^{gnd} using only the Type I grounded condition ($s = 1$). This yields $\theta_r^{\text{gnd}} = 0.19$: ground rules are perceived as far more defeasible than abstract generalizations. This is intuitive, as a specific claim like "if presentation worry occurs, then it makes Alice lose concentration" has obvious potential exceptions for this particular person, whereas the abstract generalization "if people are worried, they find it hard to concentrate" feels more universal.

Predictions. The model generates predictions for the conditions *not used* in estimation (Table 2). The key out-of-sample test is grounded Type III ($s = 2$): the model predicts 99.0% rule-targeting versus 98.8% observed (0.2 pp error), despite θ_f being estimated from abstract scenarios and θ_r^{gnd} from grounded Type I only. Since both products $\theta_f\theta_r^{\text{abs}} = 0.45$ and $\theta_f\theta_r^{\text{gnd}} = 0.16$ fall below 0.5, the model correctly predicts that rule-targeting increases from Type I to Type III in both settings.

[2] As noted in Section 3, part of the rule/categorical asymmetry is constitutive, so the direction of H1 is unsurprising.

Setting	Type (s)	Human	GPT	Claude	Gemini	Model
Rule-targeting rate (%)						
Abstract	I (1)	80.7	66.7	66.7	74.1	80.7
	II (1)	90.7	100	100	100	80.7
	III (2)	83.4	100	83.3	83.3	83.4
Generalization ($\hat{\gamma}$)						
Abstract	II (1)	0.42	0.00	0.03	0.00	—
Grounded	I (1)	0.45	0.00	0.00	0.26	—
	II (1)	0.33	0.04	0.05	0.03	—
	III (2)	0.35	0.04	0.11	0.17	—

Table 3: LLM versus human revision patterns. *Top:* Rule-targeting rates (%) in abstract scenarios (human rates averaged across Experiments 1 and 2). *Bottom:* Generalization parameter $\hat{\gamma}$ in abstract Type II ($N = 2$, $s = 1$) and grounded scenarios. The divergence lies in how broadly revisions generalize to related beliefs, and is present in both settings.

5 LLM Evaluation

Having established how humans revise beliefs, we now test whether LLMs—increasingly used as proxies for human cognition (Aher et al., 2023; Shanahan et al., 2023)—exhibit the same patterns. If LLMs revise beliefs like humans, they could serve as scalable substitutes for cognitive experiments; if not, the decomposition into rule-targeting and generalization should reveal precisely where they diverge.

Setup. We evaluated three frontier LLMs, OpenAI's GPT 5.2, Anthropic's Claude 4.5 Opus, and Google's Gemini 3 Pro, on the same scenarios used in our human experiments. Following Shanahan et al. (2023), we employed a role-play framing, instructing each LLM to imagine itself as a person holding the given beliefs who must decide how to update them in light of contradictory information. Chain-of-thought prompting was used to elicit explicit reasoning. Each model was run 60 times per scenario (540 total responses), treating each response as an independent trial to enable direct comparison with human data.[3]

Results. Table 3 compares human and LLM revision patterns. In abstract settings, LLMs show rule-targeting rates in a broadly similar range to humans, with some inconsistency across types (e.g., all three models are 100% rule-targeting on Type II but only 67–74% on Type I). The Pearson correlations between LLM and human per-scenario revision patterns are positive but weak and nonsignificant (Claude: $r = 0.42$; GPT: $r = 0.37$; Gemini: $r = 0.39$; though with only nine scenarios, these tests have limited power).

The critical divergence is in *generalization*, and it is already visible in abstract settings. In abstract Type II, where the non-contradicted rule R_2 shares an antecedent with the contradicted R_1, humans generalize at $\gamma = 0.42$—but all three LLMs revise only the contradicted rule ($\gamma \leq 0.03$). None of the LLM responses revised the categorical alongside the rule, compared to 12% of human rule-targeters.

This gap widens in grounded scenarios. Rule-targeting is near-universal for *both* humans and LLMs ($\geq 93\%$), but humans generalize substantially ($\gamma = 0.33$–0.45), revising related ground rules beyond the one directly contradicted. LLMs almost exclusively revise only the directly contradicted instance: $\gamma_{\text{GPT}} \leq 0.04$, $\gamma_{\text{Claude}} \leq 0.11$, and γ_{Gemini} is variable but consistently below the human range. The difference in the number of rules changed per response between humans and each LLM is significant across all grounded types and models (one-sided Mann–Whitney U tests, all $p < 0.05$; for GPT and Claude, all $p < 10^{-4}$).

LLMs recognize but reject explanation-based revision. Examination of the LLMs' reasoning traces reveals that the generalization failure is not a capability limitation but an

[3]Note that our comparison concerns revision of beliefs supplied *in context*, and therefore make no claim about knowledge encoded in weights, which may be organized differently.

active choice. LLMs often *acknowledge* that generalization is plausible—e.g., Claude reasons "this could extend to other sugary drinks too"—but then explicitly invoke minimality to justify local repair: "the minimal revision would be to just address R_1." Across grounded responses, Claude invoked minimality in 44% of cases, GPT in 16%, and Gemini in 19%, with all three referencing it more in grounded than abstract tasks. GPT exemplifies the pattern, reasoning that "the other rules are not directly contradicted" and "a human would not automatically generalize this one failure without further evidence." In contrast, humans do exactly that. This suggests that LLMs have internalized formal minimality norms from their training data and default to these even when their own reasoning identifies a broader revision as plausible.

6 Discussion & Conclusion

Our investigation reveals a clear decomposition of human belief revision into two orthogonal components, *what* to revise and *how broadly* to generalize, and demonstrates that, while frontier LLMs agree with humans on the first dimension, they systematically fall short on the second in both abstract and grounded contexts. We observed that people trust directly asserted categoricals more than the universality of general rules ($\theta_f > \theta_r$). In grounded scenarios, this preference *strengthens* ($\theta_r^{\text{gnd}} = 0.19$ vs. $\theta_r^{\text{abs}} = 0.55$), as specific ground rules are perceived as highly defeasible. When rule-targeters generalize, they do so partially ($\gamma \approx 0.33$–0.45 in both abstract and grounded settings), with a trimodal distribution in grounded scenarios suggesting distinct cognitive strategies rather than a uniform tendency.

Implications for LLMs as cognitive models. Our findings challenge the "silicon subjects" assumption (Aher et al., 2023; Shanahan et al., 2023) for belief revision tasks. Even in abstract scenarios, where LLMs match human rule-targeting rates, they fail to generalize: in Type II, humans propagate revisions to the related rule ($\gamma = 0.42$) and sometimes also revise the categorical (12%), while LLMs revise only the directly contradicted rule ($\gamma \leq 0.03$, 0% categorical co-revision). This aligns with the literal-versus-functional competence distinction of Riemer et al. (2025) and Mahowald et al. (2024): LLMs can simulate the *appearance* of explanation-based revision without performing the structured inference that underlies it. Importantly, our CoT analysis shows that LLMs acknowledge that generalization is plausible before invoking minimality to justify minimal revision, suggesting the gap reflects training bias toward formal revision norms rather than an inability to reason about related beliefs. The implication is that LLM-based cognitive simulations may systematically misrepresent human reasoning in domains requiring structured knowledge representations.

Limitations. Our experiments use controlled, simplified scenarios—a standard methodology in cognitive psychology for isolating mechanisms, but one that leaves open the question of generalization to naturalistic settings with ambiguous information and complex belief structures. The model assumes independent priors across rules, which fails to capture the elevated rule-targeting rate in Type II scenarios where rules share an antecedent; the model under-predicts Type II by 10 percentage points in abstract and 2.5 in grounded settings, its largest systematic error. The generalization parameter γ varies across scenario types (0.33–0.45), suggesting it may reflect scenario-specific factors rather than a stable individual trait. Finally, our LLM evaluation is a snapshot of rapidly evolving technology; future architectures may exhibit different patterns.

Future directions. The trimodal distribution of generalization strategies observed in Experiment 3 invites a hierarchical Bayesian extension with participant-level parameters, which could distinguish genuine individual differences from scenario-specific variation. Extending the model with correlated priors across rules sharing an antecedent would address the Type II gap. Our LLM evaluation also speaks only to the revision of beliefs presented in context (Section 5). Whether the knowledge encoded in a model's weights is organized so as to support propagation, and whether revision applied there through model editing or finetuning behaves more like human generalization, is an important complementary question, and one that would help separate a representational account from a task-format one.

Acknowledgments

This work was partially supported by the US Office of Naval Research under grant N00014-24-1-2663 and the National Science Foundation under award 2232055. The views and conclusions expressed in this paper are those of the authors and do not necessarily reflect the official policies or positions of the sponsoring organizations, agencies, or governments. Our special thanks go to the reviewers for their valuable feedback.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.

Carlos E Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

John R Anderson. *The Adaptive Character of Thought*. Lawrence Erlbaum Associates, 1990.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Marcel Binz and Eric Schulz. Turning large language models into cognitive models. In *International Conference on Learning Representations*, 2024.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Cem Demircan, Maria K Eckstein, Thomas L Griffiths, et al. Centaur: a foundation model of human cognition. *Nature*, 2025.

Ruth MJ Byrne and Clare R Walsh. Contradictions and counterfactuals: Generating belief revisions in conditional inference. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2002.

Renée Elio and Francis Jeffry Pelletier. Belief change as propositional update. *Cognitive Science*, 21(4):419–460, 1997.

Eduardo Fermé, Marco Garapa, Abhaya Nayak, and Maurício DL Reis. Relevance, recovery and recuperation: A prelude to ring withdrawal. *International Journal of Approximate Reasoning*, 166:109108, 2024.

Thomas L Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B Tenenbaum. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364, 2010.

Thomas L Griffiths, Jian-Qiao Zhu, Erin Grant, and R Thomas McCoy. Bayes in the age of intelligent machines. *arXiv preprint arXiv:2311.10206*, 2023.

Thomas L Griffiths, Nick Chater, and Joshua B Tenenbaum. *Bayesian Models of Cognition: Reverse Engineering the Mind*. MIT Press, 2024.

Peter Hase, Mohit Bansal, Peter Clark, and Sarah Wiegreffe. Fundamental problems with model editing: How should rational belief revision work in LLMs? *arXiv preprint arXiv:2406.19354*, 2024.

William James. Pragmatism's conception of truth. *Journal of Philosophy, Psychology and Scientific Methods*, 4(6):141–155, 1907.

Sangeet Khemlani and PN Johnson-Laird. Cognitive changes from explanations. *Journal of Cognitive Psychology*, 25(2):139–146, 2013.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540, 2024.

David Makinson. On the force of some apparent counterexamples to recovery. In E. G. Valdés, W. Krawietz, G. H. von Wright, and R. Zimmerling (eds.), *Normative Systems in Legal and Moral Theory*, pp. 475–481. 1997.

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman and Company, 1982.

Stefan Palan and Christian Schitter. Prolific: A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

Guy Politzer and Laure Carles. Belief revision and uncertain reasoning. *Thinking & Reasoning*, 7(3):217–234, 2001.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D Weisz, and Murray Campbell. Position: Theory of mind benchmarks are broken for large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Hans Rott. Two dogmas of belief revision. *Journal of Philosophy*, 97(9):503–522, 2000.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

Clare R Walsh and PN Johnson-Laird. Changing your mind. *Memory & Cognition*, 37(5): 624–631, 2009.

Bryan Wilie, Ondrej Dusek, Samuel Cahyawijaya, and Pascale Fung. Belief revision: The adaptability of large language models reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024.

Jian-Qiao Zhu, Haijiang Yan, and Thomas L Griffiths. Language models trained to do arithmetic predict human risky and intertemporal choice. In *International Conference on Learning Representations*, 2025.

	Problem Type	Non-Minimal Revision	Minimal Revision	Wilcoxon Test (p-value)	Effect Size (Cohen's d)
Experiment 1	*Type I*	132 (81.99%)	29 (18.01%)	1.14×10^{-7}	0.87
	Type II	140 (86.96%)	21 (13.04%)	8.64×10^{-11}	1.45
	Type III	144 (81.36%)	33 (18.64%)	1.25×10^{-7}	0.95
	Aggregate	416 (83.37%)	83 (16.63%)	6.41×10^{-10}	1.19
Experiment 2	*Type I*	131 (79.39%)	34 (20.61%)	3.86×10^{-8}	0.95
	Type II	155 (94.50%)	9 (5.50%)	1.69×10^{-12}	3.06
	Type III	129 (85.43%)	22 (14.57%)	1.14×10^{-8}	1.14
	Aggregate	415 (86.46%)	65 (13.54%)	4.77×10^{-11}	1.66
Experiment 3	*Type I*	115 (73.20%)	42 (26.80%)	1.72×10^{-6}	0.76
	Type II	101 (66.00%)	52 (34.00%)	2.22×10^{-3}	0.40
	Type III	107 (66.00%)	55 (34.00%)	8.08×10^{-4}	0.44
	Aggregate	323 (68.40%)	149 (31.60%)	1.98×10^{-5}	0.63

Table 4: Count-based (non-minimal vs. minimal) classification across all three experiments, with *Aggregate* representing combined data from all problem types. For Experiments 1 and 2, non-minimal coincides with rule-targeting. For Experiment 3, the count-based classification differs from the type-based rule-targeting classification reported in Table 1 of the main text: a response is non-minimal here if the *number* of belief changes exceeds the minimal threshold, whereas Table 1 classifies by *whether* at least one rule was targeted. The latter yields substantially higher rates (95–99%) because most rule-targeters revise only the directly contradicted ground rule.

A Appendix

A.1 Human Experiments

In what follows, we describe the three experiments that we undertook, and provide some additional analysis of the results.

Table 4 shows the detailed results from all experiments.

A.1.1 Experiment 1

Participants in this experiment were given the following information and instructions:

> *You will be presented with a series of everyday common scenarios. In each scenario, you will be presented with information from two or three different speakers talking about some specific things. You will then be given some additional information that you know, for a fact, to be true. Your task, in essence, is to explain what is going on.*

- ***Read carefully:*** *For each scenario, read all the information very carefully.*
- ***Explain****: Think about how to explain the fact. In other words, ask yourself: why does the fact conflict with the information provided by the speakers? Answer in your own words.*
- ***No Right or Wrong Answers****: This study aims to understand your personal thought process. There are no right or wrong answers. Choose what feels most accurate to you.*
- ***Pace Yourself****: While there's no strict time limit, try to spend a reasonable amount of time on each scenario—neither rushing through nor overthinking too much.*

Afterwards, the participants saw the following nine scenarios, and their task was to answer the corresponding question:

Scenario 1 (Type I):

- $\mathcal{R}_1$: *If a drink contains sugar, then it gives you energy.*

- $\mathcal{F}_1$: *This drink contains sugar.*
- **Fact**: *In fact, it doesn't give you energy.*

Why does the drink not give you energy?

Scenario 2 (Type I):

- $\mathcal{R}_1$: *If sales go up, then profits improve.*
- $\mathcal{F}_1$: *The sales went up.*
- **Fact**: *In fact, the profits did not go up.*

Why did the profits not go up?

Scenario 3 (Type I):

- $\mathcal{R}_1$: *If people have a fever, then they have a high temperature.*
- $\mathcal{F}_1$: *Maria had a fever.*
- **Fact**: *In fact, Maria did not have a high temperature.*

Why did Maria not have a high temperature?

Scenario 4 (Type II):

- $\mathcal{R}_1$: *If there is very loud music, then it is difficult to have a conversation.*
- $\mathcal{R}_2$: *If there is very loud music, then the neighbors complain.*
- $\mathcal{F}_1$: *The music was loud.*
- **Fact**: *In fact, the neighbors did not complain.*

Why did the neighbors not complain?

Scenario 5 (Type II):

- $\mathcal{R}_1$: *If people are worried, then they find it difficult to concentrate.*
- $\mathcal{R}_2$: *If people are worried, then they have insomnia.*
- $\mathcal{F}_1$: *Alice was worried.*
- **Fact**: *In fact, Alice did not find it difficult to concentrate.*

Why did Alice not find it difficult to concentrate?

Scenario 6 (Type II):

- $\mathcal{R}_1$: *If you follow this diet, then you lose weight.*
- $\mathcal{R}_2$: *If you follow this diet, then you have a good supply of iron*
- $\mathcal{F}_1$: *John followed this diet.*
- **Fact**: *In fact, John did not lose weight.*

Why did John not lose weight?

Scenario 7 (Type III):

- $\mathcal{R}_1$: *If someone is very kind to you, then you like that person.*
- $\mathcal{R}_2$: *If someone is very kind to you, then you are kind in return.*
- $\mathcal{F}_1$: *Jocko is very kind to Kristen.*
- **Fact**: *In fact, Kristen did not like Jocko, and she was not kind in return.*

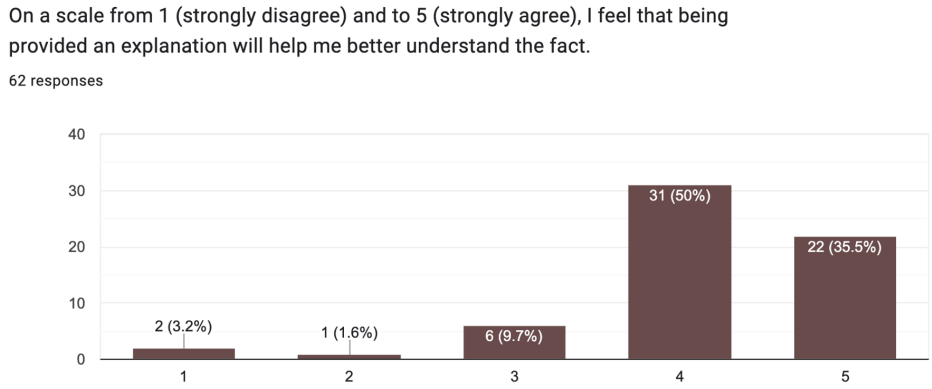


Figure 1: Distribution of responses to the Likert-type question Q2 in Experiment 1.

Why did Kristen not like Jocko and was not kind to him?

Scenario 8 (Type III):

- $\mathcal{R}_1$: *If a match is struck, then it produces light.*
- $\mathcal{R}_2$: *If a match is struck, then it gives off smoke.*
- $\mathcal{F}_1$: *Mary struck a match.*
- **Fact**: *In fact, the match produced no light, and it did not give off smoke.*

Why did the match produce no light and gave off no smoke?

Scenario 9 (Type III):

- $\mathcal{R}_1$: *If people are nervous, then their hands shake.*
- $\mathcal{R}_2$: *If people are nervous, then they get butterflies in their stomach.*
- $\mathcal{F}_1$: *Patrick was nervous.*
- **Fact**: *In fact, Patrick's hands did not shake, and he didn't get butterflies in his stomach.*

Why did Patrick's hands not shake and he didn't get butterflies in his stomach?

After going through all nine scenarios, the participants were finally asked the following two questions:

Q1: *Describe in your own words how you approached explaining what was going on. Was there a specific reason why you chose to retain or discard certain information?*

Q2: *On a scale from 1 (strongly disagree) and to 5 (strongly agree), I feel that being provided an explanation will help me better understand the fact.*

Figure 1 shows the distribution of the Likert question (Q2).

A.1.2 Experiment 2

In this experiment, the participants saw the same nine scenarios as in Experiment 1 but with a corresponding explanation that explains the inconsistency. Their task was to describe how they would revise their information in light of the given explanation.

The scenarios the participants saw can be seen below:

Scenario 1 (Type I):

- $\mathcal{R}_1$: *If a drink contains sugar, then it gives you energy.*

- $\mathcal{F}_1$: *This drink contains sugar.*
- **Fact**: *In fact, it doesn't give you energy.*
- **Explanation:** *If a person has metabolic disorders, then a sugary drink may not provide energy.*

Why does the drink not give you energy?

Scenario 2 (Type I):

- $\mathcal{R}_1$: *If sales go up, then profits improve.*
- $\mathcal{F}_1$: *The sales went up.*
- **Fact**: *In fact, the profits did not go up.*
- **Explanation**: *If expenses rise at a faster rate than sales, then an increase in sales may not lead to improved profits.*

Why did the profits not go up?

Scenario 3 (Type I):

- $\mathcal{R}_1$: *If people have a fever, then they have a high temperature.*
- $\mathcal{F}_1$: *Maria had a fever.*
- **Fact**: *In fact, Maria did not have a high temperature.*
- **Explanation**: *If a person has taken antipyretics, then they may not have a high temperature.*

Scenario 4 (Type II):

- $\mathcal{R}_1$: *If there is very loud music, then it is difficult to have a conversation.*
- $\mathcal{R}_2$: *If there is very loud music, then the neighbors complain.*
- $\mathcal{F}_1$: *The music was loud.*
- **Fact**: *In fact, the neighbors did not complain.*
- **Explanation:** *If the neighbors are away on vacations, then very loud music does not lead to complaints.*

Scenario 5 (Type II):

- $\mathcal{R}_1$: *If people are worried, then they find it difficult to concentrate.*
- $\mathcal{R}_2$: *If people are worried, then they have insomnia.*
- $\mathcal{F}_1$: *Alice was worried.*
- **Fact**: *In fact, Alice did not find it difficult to concentrate.*
- **Explanation:** *If people have effective coping strategies, then they may still be able to concentrate despite being worried.*

Scenario 6 (Type II):

- $\mathcal{R}_1$: *If you follow this diet, then you lose weight.*
- $\mathcal{R}_2$: *If you follow this diet, then you have a good supply of iron*
- $\mathcal{F}_1$: *John followed this diet.*
- **Fact**: *In fact, John did not lose weight.*
- **Explanation:** *If people have metabolic imbalances, then following a particular diet may not result in weight loss.*

Scenario 7 (Type III):

- $\mathcal{R}_1$: *If someone is very kind to you, then you like that person.*
- $\mathcal{R}_2$: *If someone is very kind to you, then you are kind in return.*
- $\mathcal{F}_1$: *Jocko is very kind to Kristen.*
- **Fact**: *In fact, Kristen did not like Jocko, and she was not kind in return.*
- **Explanation:** *If people have had negative past experiences with someone, then they may not like that person or reciprocate kindness despite the person being kind to them.*

Scenario 8 (Type III):

- $\mathcal{R}_1$: *If a match is struck, then it produces light.*
- $\mathcal{R}_2$: *If a match is struck, then it gives off smoke.*
- $\mathcal{F}_1$: *Mary struck a match.*
- **Fact**: *In fact, the match produced no light, and it did not give off smoke.*
- **Explanation:** *If the match is wet, then it will neither produce light nor give off smoke.*

Scenario 9 (Type III):

- $\mathcal{R}_1$: *If people are nervous, then their hands shake.*
- $\mathcal{R}_2$: *If people are nervous, then they get butterflies in their stomach.*
- $\mathcal{F}_1$: *Patrick was nervous.*
- **Fact**: *In fact, Patrick's hands did not shake, and he didn't get butterflies in his stomach.*
- **Explanation:** *If individuals have practiced stress-management techniques, then they may not exhibit shaky hands or butterflies in the stomach when nervous.*

After each single scenario, the participants answered the following question:

Describe in your own words how you will revise the information. Was there a specific reason why you chose to retain or discard information from the speakers? To be brief, you can write: keep $\mathcal{R}_1$, discard $\mathcal{R}_1$, alter $\mathcal{R}_1$, and so on (if you alter, please describe how).

Figure 2 plots the distribution of average number of belief changes in the nine scenarios,

A.1.3 Experiment 3

In this experiments, the participants saw the nine scenarios depicted below, and their task was to indicate which statements they will discard, alter, or keep. To better understand participants' decision-making processes, we also collected qualitative data at the end of the experiment by asking the participants three subjective questions, such as about their confidence in their revision decisions and if they considered how the explanation might apply to beliefs beyond the specific contradiction. Figure 2 plots the distribution of average number of belief changes in the nine scenarios, while Figure 3 shows the distribution of the answers to three subjective questions.

Scenario 1 (Type I):

- $\mathcal{R}_1$: *If orange juice contains sugar, then orange juice gives Tom energy.*
- $\mathcal{R}_2$: *If orange juice contains sugar, then orange juice gives Sarah energy.*
- $\mathcal{R}_3$: *If cola contains sugar, then cola gives Tom energy.*
- $\mathcal{R}_4$**:** *If cola contains sugar, then cola gives Sarah energy.*
- $\mathcal{F}_1$: *This orange juice contains sugar.*
- $\mathcal{F}_2$: *This cola contains sugar.*

- **Fact**: *In fact, the orange juice did not give Tom energy.*
- **Explanation:** *If a person has metabolic disorders, then a sugary drink may not provide energy.*

Scenario 2 (Type I):

- $\mathcal{R}_1$: *If electronics sales increase, then electronics profits improve for Store A.*
- $\mathcal{R}_2$: *If electronics sales increase, then electronics profits improve for Store B.*
- $\mathcal{R}_3$: *If clothing sales increase, then clothing profits improve for Store A.*
- $\mathcal{R}_4$: *If clothing sales increase, then clothing profits improve for Store B.*
- $\mathcal{F}_1$: *Electronics sales went up.*
- $\mathcal{F}_2$: *Clothing sales went up.*
- **Fact**: *In fact, electronics profits did not improve for Store A.*
- **Explanation**: *If expenses rise at a faster rate than sales, then an increase in sales may not lead to improved profits.*

Scenario 3 (Type I):

- $\mathcal{R}_1$: *If morning fever occurs, then morning fever causes Maria's temperature to rise.*
- $\mathcal{R}_2$: *If morning fever occurs, then morning fever causes Robert's temperature to rise.*
- $\mathcal{R}_3$: *If evening fever occurs, then evening fever causes Maria's temperature to rise.*
- $\mathcal{R}_4$: *If evening fever occurs, then evening fever causes Robert's temperature to rise.*
- $\mathcal{F}_1$: *Morning fever occurred.*
- $\mathcal{F}_2$: *Evening fever occurred.*
- **Fact**: *In fact, Maria's temperature did not rise during her morning fever.*
- **Explanation**: *If a person has taken antipyretics, then they may not have a high temperature.*

Scenario 4 (Type II):

- $\mathcal{R}_1$: *If rock music is loud, then rock music makes conversation difficult for the Browns.*
- $\mathcal{R}_2$: *If rock music is loud, then rock music makes conversation difficult for the Smiths.*
- $\mathcal{R}_3$: *If electronic music is loud, then electronic music makes conversation difficult for the Browns.*
- $\mathcal{R}_4$: *If electronic music is loud, then electronic music makes conversation difficult for the Smiths.*
- $\mathcal{R}_5$: *If rock music is loud, then rock music makes the Browns complain.*
- $\mathcal{R}_6$: *If rock music is loud, then rock music makes the Smiths complain.*
- $\mathcal{R}_7$: *If electronic music is loud, then electronic music makes the Browns complain.*
- $\mathcal{R}_8$: *If electronic music is loud, then electronic music makes the Smiths complain.*
- $\mathcal{F}_1$: *The rock music was loud.*
- $\mathcal{F}_2$: *The electronic music was loud.*
- **Fact**: *In fact, the Browns did not complain about the rock music.*
- **Explanation**: *If the neighbors are away on vacations, then very loud music does not lead to complaints.*

Scenario 5 (Type II):

- $\mathcal{R}_1$: *If presentation worry occurs, then presentation worry makes Alice lose concentration.*

- $\mathcal{R}_2$: *If presentation worry occurs, then presentation worry makes John lose concentration.*
- $\mathcal{R}_3$: *If exam worry occurs, then exam worry makes Alice lose concentration.*
- $\mathcal{R}_4$: *If exam worry occurs, then exam worry makes John lose concentration.*
- $\mathcal{R}_5$: *If presentation worry occurs, then presentation worry gives Alice insomnia.*
- $\mathcal{R}_6$: *If presentation worry occurs, then presentation worry gives John insomnia.*
- $\mathcal{R}_7$: *If exam worry occurs, then exam worry gives Alice insomnia.*
- $\mathcal{R}_8$: *If exam worry occurs, then exam worry gives John insomnia.*
- $\mathcal{F}_1$: *Presentation worry occurred.*
- $\mathcal{F}_2$: *Exam worry occurred.*
- **Fact**: *In fact, Alice did not lose concentration during her presentation.*
- **Explanation**: *If people have effective coping strategies, then they may still be able to concentrate despite being worried.*

Scenario 6 (Type II):

- $\mathcal{R}_1$: *If Mediterranean diet is followed, then Mediterranean diet helps David lose weight.*
- $\mathcal{R}_2$: *If Mediterranean diet is followed, then Mediterranean diet helps Emma lose weight.*
- $\mathcal{R}_3$: *If Keto diet is followed, then Keto diet helps David lose weight.*
- $\mathcal{R}_4$: *If Keto diet is followed, then Keto diet helps Emma lose weight.*
- $\mathcal{R}_5$: *If Mediterranean diet is followed, then Mediterranean diet gives David good iron levels.*
- $\mathcal{R}_6$: *If Mediterranean diet is followed, then Mediterranean diet gives Emma good iron levels.*
- $\mathcal{R}_7$: *If Keto diet is followed, then Keto diet gives David good iron levels.*
- $\mathcal{R}_8$: *If Keto diet is followed, then Keto diet gives Emma good iron levels.*
- $\mathcal{F}_1$: *Mediterranean diet was followed.*
- $\mathcal{F}_2$: *Keto diet was followed.*
- **Fact**: *In fact, David did not lose weight on the Mediterranean diet.*
- **Explanation**: *If people have metabolic imbalances, then following a particular diet may not result in weight loss.*

Scenario 7 (Type III):

- $\mathcal{R}_1$: *If classroom kindness occurs, then classroom kindness makes Jocko like Kristen.*
- $\mathcal{R}_2$: *If classroom kindness occurs, then classroom kindness makes Kristen like Jocko.*
- $\mathcal{R}_3$: *If office kindness occurs, then office kindness makes Jocko like Kristen.*
- $\mathcal{R}_4$: *If office kindness occurs, then office kindness makes Kristen like Jocko.*
- $\mathcal{R}_5$: *If classroom kindness occurs, then classroom kindness makes Jocko kind in return.*
- $\mathcal{R}_6$: *If classroom kindness occurs, then classroom kindness makes Kristen kind in return.*
- $\mathcal{R}_7$: *If office kindness occurs, then office kindness makes Jocko kind in return.*
- $\mathcal{R}_8$: *If office kindness occurs, then office kindness makes Kristen kind in return.*
- $\mathcal{F}_1$: *Classroom kindness occurred.*
- $\mathcal{F}_2$: *Office kindness occurred.*
- **Fact**: *In fact, Kristen did not like Jocko despite his classroom kindness, and she was not kind in return.*
- **Explanation**: *If people have had negative past experiences with someone, then they may not like that person or reciprocate kindness despite the person being kind to them.*

Scenario 8 (Type III):

- $\mathcal{R}_1$: *If wooden match is struck, then wooden match produces light for Jane.*
- $\mathcal{R}_2$: *If wooden match is struck, then wooden match produces light for Peter.*
- $\mathcal{R}_3$: *If safety match is struck, then safety match produces light for Jane.*
- $\mathcal{R}_4$: *If safety match is struck, then safety match produces light for Peter.*
- $\mathcal{R}_5$: *If wooden match is struck, then wooden match gives off smoke for Jane.*
- $\mathcal{R}_6$: *If wooden match is struck, then wooden match gives off smoke for Peter.*
- $\mathcal{R}_7$: *If safety match is struck, then safety match gives off smoke for Jane.*
- $\mathcal{R}_8$: *If safety match is struck, then safety match gives off smoke for Peter.*
- $\mathcal{F}_1$: *Wooden match was struck.*
- $\mathcal{F}_2$: *Safety match was struck.*
- **Fact**: *In fact, the wooden match did not produce light or smoke for Jane.*
- **Explanation**: *If a match is wet, then it will neither produce light nor give off smoke.*

Scenario 9 (Type III):

- $\mathcal{R}_1$: *If speech nervousness occurs, then speech nervousness makes Patrick's hands shake.*
- $\mathcal{R}_2$: *If speech nervousness occurs, then speech nervousness makes Diana's hands shake.*
- $\mathcal{R}_3$: *If interview nervousness occurs, then interview nervousness makes Patrick's hands shake.*
- $\mathcal{R}_4$: *If interview nervousness occurs, then interview nervousness makes Diana's hands shake.*
- $\mathcal{R}_5$: *If speech nervousness occurs, then speech nervousness gives Patrick butterflies.*
- $\mathcal{R}_6$: *If speech nervousness occurs, then speech nervousness gives Diana butterflies.*
- $\mathcal{R}_7$: *If interview nervousness occurs, then interview nervousness gives Patrick butterflies.*
- $\mathcal{R}_8$: *If interview nervousness occurs, then interview nervousness gives Diana butterflies.*
- $\mathcal{F}_1$: *Speech nervousness occurred.*
- $\mathcal{F}_2$: *Interview nervousness occurred.*
- **Fact**: *In fact, Patrick's hands did not shake during his speech and he didn't get butterflies.*
- **Explanation**: *If individuals have practiced stress-management techniques, then they may not exhibit shaky hands or butterflies in the stomach when nervous.*

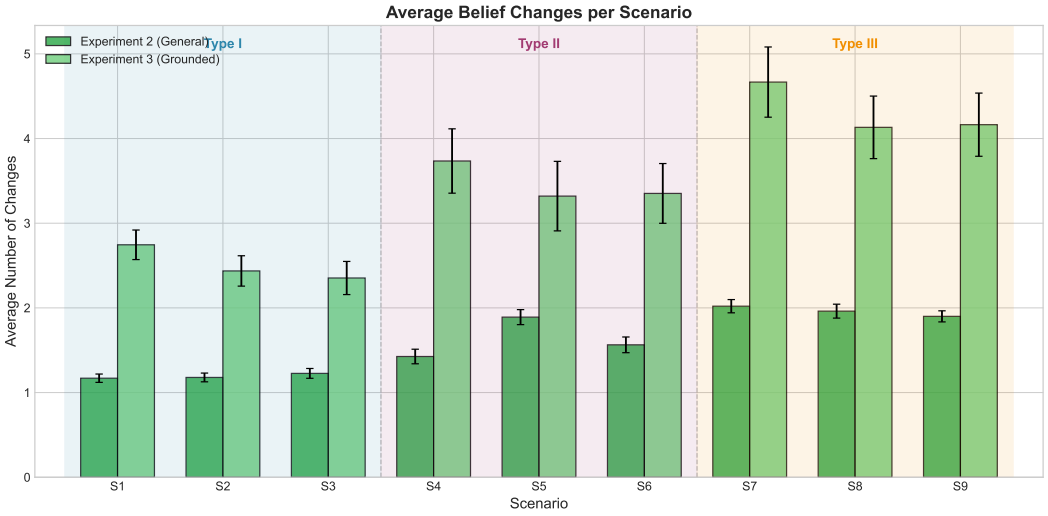
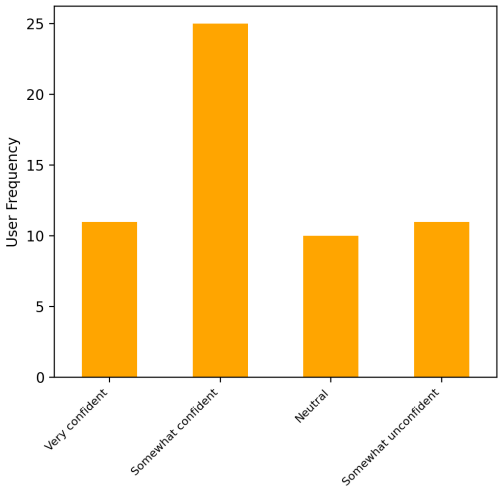
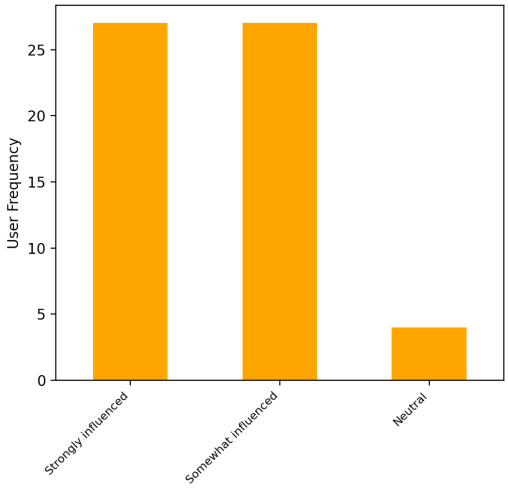


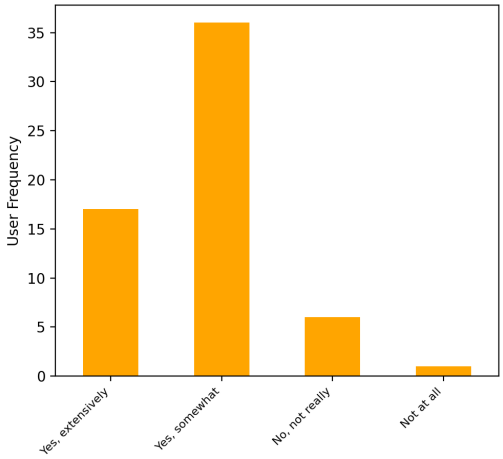
Figure 2: Average number of belief changes per scenario in Experiments 2 and 3.



(a) Confidence in revision decisions



(b) Influence of explanations



(c) Consideration of broader applications

Figure 3: Qualitative feedback from participants in Experiment 3 showing (a) their confidence levels in revision decisions, (b) the extent to which explanations influenced their decisions, and (c) whether they considered how explanations might apply beyond specific contradictions.

A.2 Bayesian Model Details

A.2.1 General Derivation for Arbitrary s and q

We derive the general rule-targeting probability for s contradicted rules and q involved categoricals under the fully crossed assumption: every contradicted rule R_i ($i = 1, \ldots, s$) paired with every involved categorical F_j ($j = 1, \ldots, q$) produces a prediction that φ contradicts. Each $U_i \sim \text{Bernoulli}(\theta_r)$ and $V_j \sim \text{Bernoulli}(\theta_f)$ independently. A joint state $(\mathbf{U}, \mathbf{V})$ survives conditioning on φ if and only if, for every pair (i, j), either $U_i = 0$ (rule defeasible) or $V_j = 0$ (categorical incorrect).

The surviving states partition as follows:

- *All rules defeasible* ($U_i = 0$ for all i): any categorical configuration is consistent. Mass: $(1-\theta_r)^s$.
- *At least one rule universal* ($\exists\, i : U_i = 1$): then for every involved F_j, the constraint requires $V_j = 0$. So all q categoricals must be incorrect. Mass: $\left[1-(1-\theta_r)^s\right](1-\theta_f)^q$.

The normalizing constant is

$$Z = (1-\theta_r)^s + \left[1-(1-\theta_r)^s\right](1-\theta_f)^q = 1 - \left[1-(1-\theta_r)^s\right]\left[1-(1-\theta_f)^q\right],$$

where the second form follows by inclusion-exclusion. The categorical-only revision probability (all rules universal, all categoricals incorrect) has mass $\theta_r^s(1-\theta_f)^q$, giving

$$P(\text{rule} \mid \varphi) = 1 - \frac{\theta_r^s\,(1-\theta_f)^q}{(1-\theta_r)^s + [1-(1-\theta_r)^s](1-\theta_f)^q}.$$

The three-way posterior is:

$$\begin{aligned} P(\text{rule only} \mid \varphi) &= \frac{(1-\theta_r)^s\,\theta_f^q}{Z}, \\ P(\text{cat. only} \mid \varphi) &= \frac{\theta_r^s\,(1-\theta_f)^q}{Z}, \\ P(\text{both} \mid \varphi) &= 1 - P(\text{rule only}) - P(\text{cat. only}). \end{aligned}$$

Reduction to $q = 1$. When $q = 1$, the normalizing constant becomes $Z = (1-\theta_r)^s + [1-(1-\theta_r)^s](1-\theta_f) = \theta_f(1-\theta_r)^s + (1-\theta_f)$, and the rule-targeting probability simplifies to Equation 4 with $q = 1$:

$$P(\text{rule} \mid \varphi) = 1 - \frac{(1-\theta_f)\,\theta_r^s}{(1-\theta_f) + \theta_f(1-\theta_r)^s},$$

which is the form used throughout the paper, since all experimental scenarios involve a single categorical in each contradiction.

A.2.2 Monotonicity Condition

The following results are stated for $q = 1$, which holds in all experimental conditions studied in this paper.

Proposition 1. *$P(\text{rule} \mid \varphi)$ is higher for $s = 2$ than for $s = 1$ if and only if $\theta_f\theta_r < 0.5$.*

Proof. From Equation 4, let $Q(s) = 1 - P(\text{rule} \mid \varphi)$ denote the categorical-only probability. P increases from s=1 to s=2 iff $Q(1) > Q(2)$. Cross-multiplying the two fractions and cancelling $(1-\theta_f)\theta_r > 0$, the inequality $Q(1) > Q(2)$ reduces to:

$$(1-\theta_r)\left[1 - 2\,\theta_f\theta_r\right] > 0.$$

Since $0 < \theta_r < 1$, the factor $(1-\theta_r)$ is positive, so the condition becomes $\theta_f\theta_r < 0.5$. □

	θ_r=0.2	θ_r=0.3	θ_r=0.4	θ_r=0.5	θ_r=0.6	θ_r=0.7	θ_r=0.8
θ_f=0.5	88.9	82.4	75.0	66.7	57.1	46.2	33.3
θ_f=0.6	90.9	85.4	78.9	71.4	62.5	51.7	38.5
θ_f=0.7	93.0	88.6	83.3	76.9	69.0	58.8	45.5
θ_f=0.8	**95.2**	92.1	88.2	**83.3**	76.9	68.2	55.6
θ_f=0.9	97.6	95.9	93.8	90.9	87.0	81.1	71.4

Table 5: $P(\text{rule} \mid \varphi)$ (%) for $s = 1$. Bold entries correspond to the fitted parameter regions: ($\theta_f \approx 0.8, \theta_r \approx 0.2$) for the grounded setting and ($\theta_f \approx 0.8, \theta_r \approx 0.5$) for the abstract setting.

	θ_r=0.2	θ_r=0.3	θ_r=0.4	θ_r=0.5	θ_r=0.6	θ_r=0.7	θ_r=0.8
θ_f=0.5	97.6	94.0	88.2	80.0	69.0	55.0	38.5
θ_f=0.6	98.0	94.8	89.6	81.8	71.0	56.8	39.6
θ_f=0.7	98.4	95.8	91.3	84.2	73.8	59.5	41.5
θ_f=0.8	**98.9**	97.0	93.4	**87.5**	78.0	64.0	44.8
θ_f=0.9	99.4	98.3	96.2	92.3	85.2	72.9	52.9

Table 6: $P(\text{rule} \mid \varphi)$ (%) for $s = 2$. Bold entries as in Table 5.

A.2.3 Categorical Co-revision Rate

For $s = 1$ and $q = 1$, the probability that a rule-targeter also revises the categorical simplifies to a function of θ_f alone. From the three-way posterior (Equations 1–3):

$$P(\text{rule \& cat.} \mid \text{rule}, \varphi) = \frac{P(\text{rule \& cat.} \mid \varphi)}{P(\text{rule} \mid \varphi)} = \frac{(1-\theta_f)(1-\theta_r)/Z}{\left[\theta_f(1-\theta_r) + (1-\theta_f)(1-\theta_r)\right]/Z}$$
$$= \frac{1-\theta_f}{\theta_f + (1-\theta_f)} = 1 - \theta_f$$

The $(1-\theta_r)$ terms cancel, so this rate depends only on how much agents trust their categoricals. With $\theta_f = 0.81$, the model predicts 19% of rule-targeters also revise the categorical.

A.2.4 Parameter Sensitivity

Tables 5 and 6 show $P(\text{rule} \mid \varphi)$ as a percentage for a grid of (θ_f, θ_r) values at $s = 1$ and $s = 2$, respectively. The fitted parameter region is highlighted: $\theta_f \approx 0.81$ (closest row: $\theta_f = 0.8$) with $\theta_r^{\text{abs}} \approx 0.55$ (closest column: $\theta_r = 0.5$) for the abstract setting and $\theta_r^{\text{gnd}} \approx 0.19$ (closest column: $\theta_r = 0.2$) for the grounded setting.

A.2.5 Model Fitting Details

Parameters were estimated via grid search over $(\theta_f, \theta_r) \in (0,1)^2$ at 0.0001 resolution, minimizing $\sum_t (\hat{P}_t - P_{\text{obs},t})^2$ across the fitting conditions. The exact fitted values are $\theta_f = 0.8056$, $\theta_r^{\text{abs}} = 0.5516$ (rounded to 0.81 and 0.55), and $\theta_r^{\text{gnd}} = 0.1951$ (rounded to 0.19). Table 7 compares predictions from exact vs. rounded parameters.

A.3 Generalization Analysis

A.3.1 Distribution of Rules Changed

Table 8 shows the distribution of the number of ground rules changed across the three problem types in Experiment 3. For Type I, the pool size is $N = 4$ ground rules with $s = 1$ rule directly contradicted. For Types II and III, the pool size is $N = 8$ ground rules.

The distribution for Type I is trimodal: roughly equal proportions of participants perform local repair ($\gamma_i = 0$; 34%), partial generalization (30%), or full theory-level revision ($\gamma_i = 1$;

Setting	Type	s	Observed	Exact	Rounded	Δ(exact)	Δ(rounded)
Abstract	I	1	80.7%	80.7%	81.2%	0.0	0.5
Abstract	II	1	90.8%	80.7%	81.2%	10.1	9.6
Abstract	III	2	83.4%	83.4%	83.8%	0.0	0.4
Grounded	I	1	95.5%	95.5%	95.7%	0.0	0.2
Grounded	II	1	98.0%	95.5%	95.7%	2.5	2.3
Grounded	III	2	98.8%	99.0%	99.1%	0.2	0.3

Table 7: Model predictions vs. observed rule-targeting rates. Types I and III (abstract) and Type I (grounded) are fitting conditions; Type III (grounded) is the key out-of-sample prediction. The Type II over-prediction reflects the independence assumption (Section 6).

Type I (N=4, s=1)			**Type II** (N=8, s=1)			**Type III** (N=8, s=2)		
# Rules	Count	%	# Rules	Count	%	# Rules	Count	%
1	51	34.0	1	54	36.0	1	15	9.4
2	45	30.0	2	24	16.0	2	57	35.6
3	6	4.0	3	7	4.7	3	12	7.5
4	48	32.0	4	30	20.0	4	25	15.6
			5	7	4.7	5	4	2.5
			6	5	3.3	6	5	3.1
			8	23	15.3	8	42	26.2

Table 8: Distribution of the number of ground rules changed per response in Experiment 3, by problem type.

Scenario	Type	s	N	Human γ	GPT γ	Claude γ	Gemini γ
S1	I	1	4	0.45	0.00	0.00	0.56
S2	I	1	4	0.45	0.00	0.00	0.17
S3	I	1	4	0.45	0.00	0.00	0.27
S4	II	1	8	0.33	0.00	0.00	0.00
S5	II	1	8	0.33	0.12	0.14	0.02
S6	II	1	8	0.33	0.00	0.00	0.09
S7	III	2	8	0.35	0.00	0.00	0.04
S8	III	2	8	0.35	0.00	0.00	0.33
S9	III	2	8	0.35	0.13	0.33	0.11

Table 9: Per-scenario generalization index γ for humans (type-level estimate) and each LLM (scenario-specific).

32%). This suggests distinct cognitive strategies rather than a single noisy tendency. For Types II and III, the distributions are more spread due to the larger statement pool ($N = 8$), but still show clustering at the extremes.

A.3.2 Per-Scenario LLM Generalization

Table 9 reports the generalization index γ for each scenario and each LLM, alongside the human estimate. The human γ column shows the type-level estimate (γ is estimated per type, not per scenario, from the aggregate data). The LLM γ values are scenario-specific.

GPT and Claude show $\gamma = 0$ in 7 out of 9 scenarios, with small non-zero values only in S5 and S9. Gemini shows more variability but remains consistently below the human range.

A.4 Scenario Robustness

Two of the nine scenarios invite objections on the grounds of construction rather than of measurement. In S3 the rule "if people have a fever, then they have a high temperature" can be read as definitional, in which case it is not clear that revising it counts as revising an empirical generalization. In S7 the two effects contradicted in the Type III condition, liking

Experiment	**Type I** (S3 flagged) Baseline	Drop S3	**Type III** (S7 flagged) Baseline	Drop S7
Exp. 1 (abstract)	82.0%	81.5%	81.4%	80.5%
Exp. 2 (abstract)	79.4%	79.5%	85.4%	84.2%
Exp. 3 (grounded)	95.5%	96.2%	98.8%	98.1%

Table 10: Human rule-targeting rates with and without the flagged scenarios. S3 appears in Type I, S7 in Type III.

Filter	θ_f	θ_r^{abs}	θ_r^{gnd}	**Grounded Type III** Obs. / Pred.	Δ
Baseline	0.806	0.552	0.194	98.8% / 99.0%	0.22
Drop S3	0.801	0.549	0.165	98.8% / 99.3%	0.52
Drop S7	0.820	0.570	0.206	98.1% / 98.9%	0.76
Drop S3 + S7	0.815	0.568	0.175	98.1% / 99.2%	1.09

Table 11: Bayesian model refitted under each exclusion. Grounded Type III is the out-of-sample condition: it is not used in estimation under any filter. Δ is the absolute error in percentage points.

someone and behaving kindly in return, are causally separable, whereas the corresponding effects in S8 (light and smoke from a match) and S9 (trembling hands and butterflies) are co-effects of a single underlying mechanism, so that a disabling condition defeating one nearly forces the other. The three Type III scenarios are therefore not perfectly matched in structure.

We re-ran the full analysis excluding each flagged scenario, and both together. This section reports the results; the per-scenario LLM estimates are in Table 9.

A.4.1 Rule-Targeting Rates Under Exclusion

Table 10 gives human rule-targeting rates for the affected conditions with and without the flagged scenario. S3 belongs to Type I and S7 to Type III, so each exclusion bears on one condition. No rate moves by more than 1.2 percentage points, and the direction of movement is not consistent across experiments: dropping S3 slightly *raises* the grounded Type I rate.

The human generalization estimates are likewise stable. For grounded Type I, γ moves from 0.447 to 0.477 when S3 is excluded; for grounded Type III, from 0.345 to 0.322 when S7 is excluded. Neither change affects the comparison with the LLMs, whose estimates lie an order of magnitude lower.

A.4.2 Refitting the Bayesian Model

Because the exclusions change the observed rates used for parameter estimation, we refit the model under each filter following the same two-step procedure as Section 4.5. Table 11 reports the fitted parameters and the key out-of-sample prediction, grounded Type III. The parameters are stable across all four variants ($\theta_f \in [0.80, 0.82]$, $\theta_r^{\text{abs}} \in [0.55, 0.57]$, $\theta_r^{\text{gnd}} \in [0.17, 0.21]$), and the out-of-sample prediction remains accurate to within about one percentage point throughout. In particular, the qualitative conclusion that grounded rules are perceived as far more defeasible than abstract ones ($\theta_r^{\text{gnd}} \ll \theta_r^{\text{abs}}$) does not depend on either scenario.

A.4.3 The Human–LLM Divergence Under Exclusion

The concern behind S7 is that its causally separable effects might inflate the apparent human tendency to generalize in Type III. Excluding it in fact raises γ for every LLM as well, and

Model	Baseline γ	Drop S7 γ	Human γ
GPT 5.2	0.04	0.07	0.35
Claude 4.5 Opus	0.11	0.17	0.35
Gemini 3 Pro	0.17	0.23	0.35

Table 12: Grounded Type III generalization with and without S7. The human estimate shown is the baseline value; excluding S7 moves it to 0.322.

leaves all three far below the human estimate (Table 12). The central divergence is therefore unaffected by the flagged scenarios.

A.5 LLM Experiments

The structure of the prompt we used for the LLM experiments can be seen below (specified for Scenario 5).

Prompt Example

Imagine you are a person who holds the following beliefs about the world. You have just learned something new that conflicts with what you believed. Think about how you would naturally update your beliefs in light of this new information. Reason the way an ordinary person would — based on what makes intuitive sense to you.

Scenario Details:

1. **Your Current Beliefs:** A set of beliefs (facts and rules) you currently hold.
2. **A New Observation:** A new piece of information you have just observed, which is a fact.

Your Current Beliefs:

R_1: If people are worried, then they find it difficult to concentrate.
R_2: If people are worried, then they have insomnia.
F_1: Alice was worried.

New Observation (Fact): In fact, Alice did not find it difficult to concentrate.

Your Task:

Part 1: Step-by-Step Reasoning First, think step-by-step about the conflict between the new observation and your current beliefs. What makes intuitive sense as an explanation?

Part 2: Final Revision Decision Based on your reasoning, provide your revision decision for each belief. Use the following structured format. For each of the original beliefs, state whether you will `keep`, `discard`, or `alter` it. If you alter a belief, you must provide the new, altered version of the rule.

R_1: $\rightarrow$ [`Keep/Discard/Alter`]
R_2: $\rightarrow$ [`Keep/Discard/Alter`]
F_1: $\rightarrow$ [`Keep/Discard/Alter`]

Altered Rules (if any):

R_{new}: [Provide the new rule here if altered]

A.5.1 One-Shot Demonstration

Section 5 characterizes LLM behavior under standard prompting. A natural question is whether the absence of generalization reflects a limit on what the models can do, or only on what they do by default. To separate these, we re-ran the evaluation with a single worked example prepended to the prompt. The example presents a contradiction to one rule and propagates the revision to a second rule sharing the same antecedent, keeping the

Setting	Type	GPT 5.2 Baseline	GPT 5.2 One-shot	Claude 4.5 Opus Baseline	Claude 4.5 Opus One-shot	Human
Abstract	II	0.00	0.10	0.03	0.16	0.42
Grounded	I	0.00	0.11	0.00	0.14	0.45
Grounded	II	0.04	0.11	0.05	0.16	0.33
Grounded	III	0.04	0.12	0.11	0.20	0.35

Table 13: Generalization γ under standard prompting (baseline) and with a single worked demonstration of a generalizing revision (one-shot). Abstract Types I and III are omitted because γ is undefined there ($N = s$).

Problem Type	Human Revisions Non-Minimal	Human Revisions Minimal	GPT 5.2 Revisions Non-Minimal	GPT 5.2 Revisions Minimal	Chi-Square (p-value)	Effect Size (Cramer's V)
Abstract Task						
Type I	131 (79.39%)	34 (20.61%)	120 (66.67%)	60 (33.33%)	0.0113653	0.14
Type II	155 (94.51%)	9 (5.49%)	180 (100.00%)	0 (0.00%)	0.00113077	0.15
Type III	129 (85.43%)	22 (14.57%)	180 (100.00%)	0 (0.00%)	1.30×10^{-8}	0.28
Aggregate	415 (86.46%)	65 (13.54%)	480 (88.89%)	60 (11.11%)	0.277518	0.03
Grounded Task						
Type I	115 (73.25%)	42 (26.75%)	0 (0.00%)	180 (100.00%)	7.14×10^{-55}	0.76
Type II	101 (66.01%)	52 (33.99%)	6 (3.33%)	174 (96.67%)	1.00×10^{-46}	0.71
Type III	107 (66.05%)	55 (33.95%)	12 (6.67%)	168 (93.33%)	3.23×10^{-34}	0.66
Aggregate	323 (68.43%)	149 (31.57%)	18 (3.33%)	522 (96.67%)	1.74×10^{-114}	0.71

Table 14: Comparison of revision patterns between Humans and GPT 5.2.

categorical, together with the reasoning that licenses the propagation. Its content (plant growth) is disjoint from all nine test scenarios, and the scenarios, response format, and parsing were otherwise unchanged. We ran GPT and Claude in this condition; Gemini was omitted because of budget constraints.

Table 13 reports the resulting estimates. The demonstration lifts γ off the floor for both models in every condition, but only partially: the largest value reached is 0.20 (Claude, grounded Type III), against human estimates of 0.33–0.45. The gap narrows and does not close.

We read this in two parts. First, human-like generalization is partly elicitable, so the baseline result is not a hard ceiling on what these models can represent; the relevant rules can be linked when the task makes the linkage salient. Second, and more importantly for our purposes, the behavior is not deployed spontaneously, and an explicit demonstration does not bring it into the human range. The silicon-subjects methodology depends on models producing human-like behavior without first being shown the behavior to produce. That an explicit demonstration is required, and still falls short, is therefore itself the relevant finding rather than a qualification of it.

More Experimental Results

The following tables compare human and LLM revision patterns using a count-based classification (number of belief changes exceeding a minimal threshold), complementing the generalization analysis in the main text.

We now present some more results from the experiments. Figure 4 shows the correlation plots per scenario, while Figure 5 shows the distribution of belief changes. Moreover, Tables 14, 15, and 16 show the statistics from comparing the LLM results and the human results from the user studies.

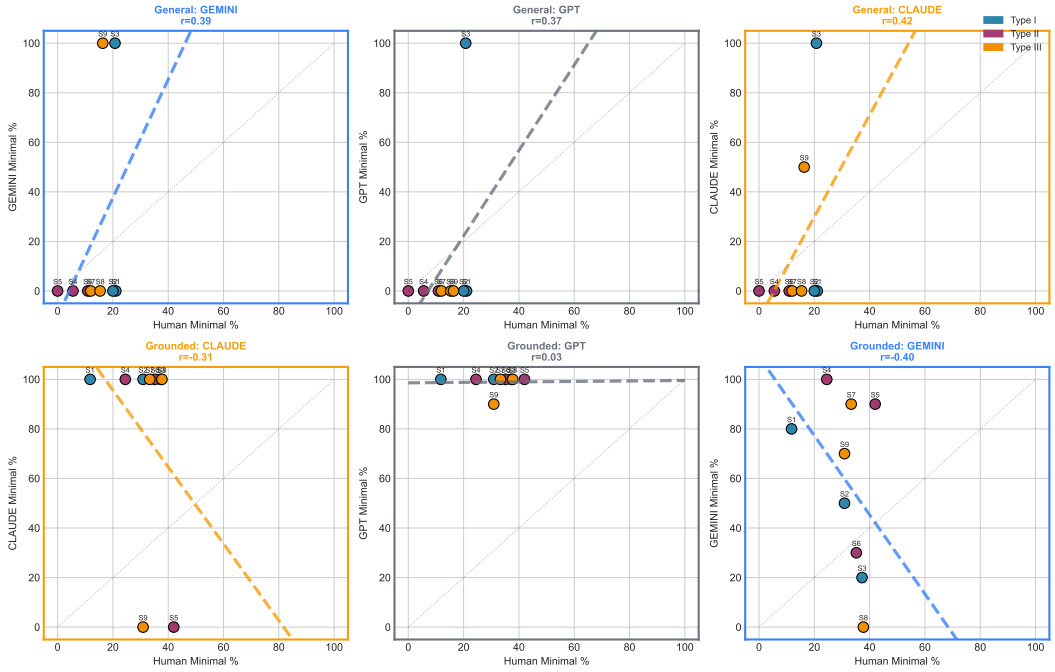


Figure 4: Detailed correlation plots of LLM vs. Human revision patterns across the nine scenarios.

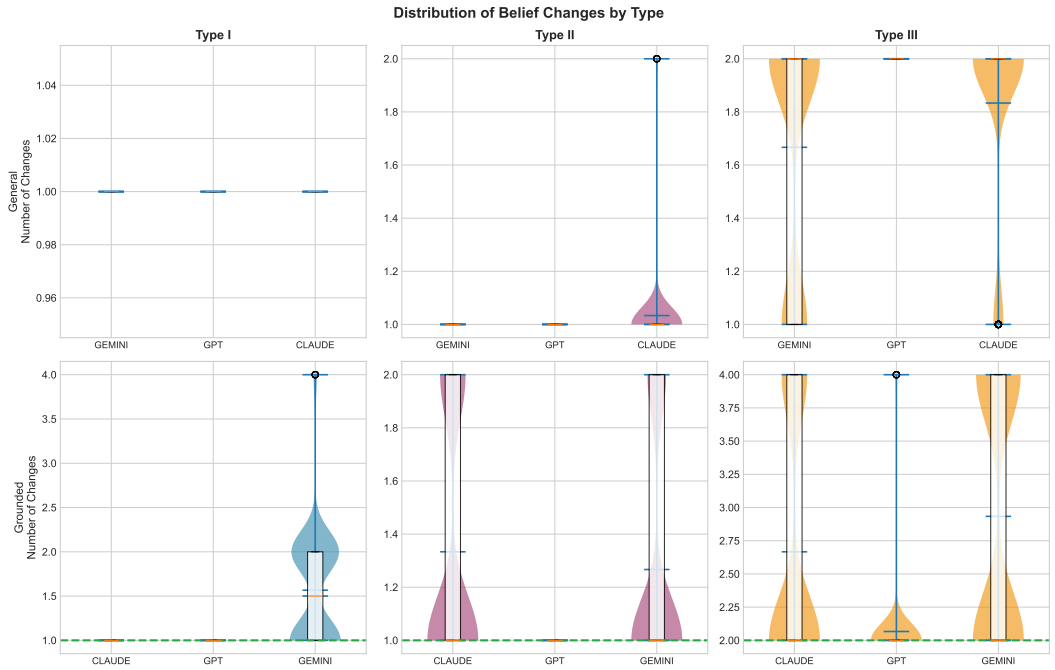


Figure 5: Distribution of average number of belief changes by Type for each LLM.

Problem Type	Human Revisions Non-Minimal	Human Revisions Minimal	Claude 4.5 Opus Revisions Non-Minimal	Claude 4.5 Opus Revisions Minimal	Chi-Square (p-value)	Effect Size (Cramer's V)
Abstract Task						
Type I	131 (79.39%)	34 (20.61%)	120 (66.67%)	60 (33.33%)	0.0113653	0.14
Type II	155 (94.51%)	9 (5.49%)	180 (100.00%)	0 (0.00%)	0.00113077	0.15
Type III	129 (85.43%)	22 (14.57%)	150 (83.33%)	30 (16.67%)	0.710936	0.02
Aggregate	415 (86.46%)	65 (13.54%)	450 (83.33%)	90 (16.67%)	0.193491	0.04
Grounded Task						
Type I	115 (73.25%)	42 (26.75%)	0 (0.00%)	180 (100.00%)	7.14×10^{-55}	0.76
Type II	101 (66.01%)	52 (33.99%)	60 (33.33%)	120 (66.67%)	5.31×10^{-9}	0.32
Type III	107 (66.05%)	55 (33.95%)	60 (33.33%)	120 (66.67%)	2.93×10^{-9}	0.32
Aggregate	323 (68.43%)	149 (31.57%)	120 (22.22%)	420 (77.78%)	4.90×10^{-49}	0.46

Table 15: Comparison of revision patterns between Humans and Claude 4.5 Opus.

Problem Type	Human Revisions Non-Minimal	Human Revisions Minimal	Gemini 3 Pro (Preview) Revisions Non-Minimal	Gemini 3 Pro (Preview) Revisions Minimal	Chi-Square (p-value)	Effect Size (Cramer's V)
Abstract Task						
Type I	131 (79.39%)	34 (20.61%)	120 (74.07%)	42 (25.93%)	0.255	0.06
Type II	155 (94.51%)	9 (5.49%)	180 (100.00%)	0 (0.00%)	0.00113077	0.15
Type III	129 (85.43%)	22 (14.57%)	120 (83.33%)	24 (16.67%)	0.620	0.03
Aggregate	415 (86.46%)	65 (13.54%)	420 (86.42%)	66 (13.58%)	0.986	0.00
Grounded Task						
Type I	115 (73.25%)	42 (26.75%)	96 (53.33%)	84 (46.67%)	2.14×10^{-5}	0.23
Type II	101 (66.01%)	52 (33.99%)	24 (13.33%)	156 (86.67%)	1.38×10^{-12}	0.39
Type III	107 (66.05%)	55 (33.95%)	84 (46.67%)	96 (53.33%)	4.74×10^{-4}	0.19
Aggregate	323 (68.43%)	149 (31.57%)	204 (37.78%)	336 (62.22%)	5.90×10^{-18}	0.27

Table 16: Comparison of revision patterns between Humans and Gemini 3 Pro (Preview).