When Humans Revise Their Beliefs, Explanations Matter: Evidence from User Studies and What It Means for AI Alignment

Stylianos Loukas Vasileiou[1], **Antonio Rago**[2], **Maria Vanina Martinez**[3], **William Yeoh**[4]
[1]New Mexico State University
[2]King's College London
[3]Artificial Intelligence Research Institute (IIIA-CSIC)
[4]Washington University in St. Louis
stelios@nmsu.edu, antonio.rago@kcl.ac.uk, vmartinez@iiia.csic.es, wyeoh@wustl.edu

Abstract

Understanding how humans revise their beliefs in light of new information is crucial for developing AI agents which can effectively model, and thus align with, human reasoning and decision-making. Motivated by empirical evidence from cognitive psychology, in this paper we first present three comprehensive human-subject studies showing that people consistently prefer explanation-based revisions, i.e., those which are guided by explanations, that result in changes to their belief agents that are more extensive than necessary. Our experiments systematically investigate how people revise their beliefs with explanations for inconsistencies, whether they are provided with them or left to formulate them themselves, demonstrating a robust preference for what may seem non-minimal revisions across different types of scenarios. Moreover, we evaluate to what extent large language models can simulate human belief revision patterns by testing state-of-the-art models on parallel tasks, analyzing their revision choices and alignment with human preferences. These findings have implications for AI agents designed to model and interact with humans, suggesting that such agents should accommodate explanation-based, potentially non-minimal belief revision operators to better align with human cognitive processes.

Introduction

For AI agents to collaborate effectively and safely with people, they must be human-aware, that is, capable of reasoning about the mental states of their human partners (Kambhampati 2020). A cornerstone of this paradigm is the AI agent's ability to build and maintain an approximate *human model* that captures its understanding of the user's beliefs, goals, and knowledge about a shared task. Such models are critical for enabling fluid, predictable, and successful interactions, whether in decision-support agents, collaborative robotics, or personalized tutoring agents.

Formally, the human-aware AI framework represents this through a multi-model setting $\mathcal{M} = \langle \mathcal{M}^R, \mathcal{M}^H \rangle$, where $\mathcal{M}^R$ denotes the AI agent's own model of the world and $\mathcal{M}^H$ represents the human model that the agent maintains (Sreedharan 2023).[1] This framework has been extensively developed to support various forms of human-AI interaction, such as the model reconciliation problem (Chakraborti et al. 2017; Vasileiou et al. 2022), where the goal is to align the models of the AI agent and the human with explanations.

A fundamental challenge in human-aware AI, however, is that these human models are incomplete and will inevitably face inconsistencies. An AI agent will frequently observe a user taking an action or stating a fact that contradicts the model's current representation of their beliefs. This triggers the problem of belief revision: the AI agent must update its model of the user to accommodate this new, conflicting information. But what principles should govern how an AI agent updates its beliefs about a human?

The predominant approach in human-aware AI literature has been to adopt the principle of minimalism (or information economy) from classical belief revision theory, which advocates making the smallest possible change to restore logical consistency (James 1907; Makinson 1997; Rott 2000; Fermé et al. 2024). This principle manifests in various algorithmic contributions that seek to minimize the number of changes to the human model $\mathcal{M}^H$ when reconciling inconsistencies (Sreedharan, Chakraborti, and Kambhampati 2021; Vasileiou, Previti, and Yeoh 2021; Son et al. 2021). The underlying assumption is that minimal changes preserve as much of the original model as possible, thereby maintaining computational efficiency as well as theoretical elegance.

However, this minimalist approach may not align with the actual cognitive processes of the human being modeled. Cognitive science research indicates that when people themselves encounter inconsistencies, they do not simply make minimal adjustments; instead, they seek to understand the conflict by generating explanations (Walsh and Johnson-Laird 2009; Khemlani and Johnson-Laird 2013). This search for explanatory understanding often leads to broader, seemingly non-minimal revisions, such as modifying relevant general rules rather than specific facts.

To illustrate this distinction, consider an AI agent with a human model consisting of the rule "If people are worried, then they find it difficult to concentrate" and the fact "Alice was worried". If the AI agent then observes new, conflicting information that "In fact, Alice did not find it difficult to concentrate", a minimalist revision might simply discard the initial fact, "Alice was worried," as this is the most localized change to restore logical consistency. However, humans are

[1] These models typically encode knowledge pertaining to a specific task, such as a planning problem (Chakraborti et al. 2017).

more likely to ask why the conflict occurred and generate an explanation, such as, “Perhaps Alice has effective coping strategies”. This explanation then guides a broader revision to accommodate this new understanding.

This creates a potential disconnect: an AI agent that adheres to minimalism may develop a brittle and inaccurate model that diverges from the human’s expectations, undermining the very goal of human-aware interaction and potentially leading to eroded trust or unsafe suggestions based on a flawed understanding of the human’s beliefs. This mismatch presents a challenge in developing human-aware AI agents.

Therefore, our goal in this paper is to empirically explore how people revise their beliefs to inform the design of more “cognitively-aligned” revision mechanisms for human-aware AI agents. We argue that for an AI agent to maintain a robust and useful model of its human user, its model-revision process should reflect the explanatory, non-minimal patterns observed in human cognition. Our results from three comprehensive user studies reveal a strong and consistent preference for revisions guided by explanations, namely *explanation-based revisions*. Furthermore, we evaluate to what extent state-of-the-art large language models (LLMs) can simulate these human belief revision patterns, providing a baseline for their current alignment with human cognitive processes. These findings can help to better understand how to design AI agents that can more accurately model their human users.

Related Work

The problem of belief revision, that is how an agent should update its knowledge in the face of new, contradictory information, is a cornerstone of AI. A dominant principle for formalizing this process has been the principle of minimalism (James 1907), which posits that a rational agent should make the smallest possible change to their beliefs to restore consistency after encountering a contradiction. However, a stream of research in cognitive science has consistently challenged the descriptive accuracy of minimalism, arguing that human reasoning is better described by an explanatory hypothesis.

This view holds that the primary goal for humans is not minimal change, but achieving explanatory understanding. Foundational studies by cognitive scientists have provided qualitative evidence that people prefer non-minimal, explanation-based revisions, often by modifying general rules to account for “disabling conditions” that explain a conflict (Elio and Pelletier 1997; Walsh and Johnson-Laird 2009; Khemlani and Johnson-Laird 2013). Our work builds directly on this foundation by providing more evidence for this human preference across a range of scenarios and, crucially, is the first to directly contrast this pattern with the revision strategies employed by state-of-the-art AI models.

Within the human-aware AI paradigm (Sreedharan 2023), a central challenge is for an AI agent to build and maintain an accurate model of its human partner, denoted $\mathcal{M}^H$. Much of the work in that space has focused on the model reconciliation problem, where the AI agent generates an explanation to align the human’s model $\mathcal{M}^H$ with its own model ($\mathcal{M}^R$), thereby making the AI agent’s behavior understandable. This research has been vital for creating more interactive, user-centered agents. Our work, however, addresses the inverse but equally critical problem: how an AI should approach revising its model of the human in response to an observation that contradicts that model. While prior approaches have often implicitly adopted minimalist updates for this task (Chakraborti et al. 2017; Vasileiou, Previti, and Yeoh 2021; Vasileiou et al. 2022), we provide an empirical demonstration of how this assumption might fail to align with actual human cognitive processes.

Finally, the evaluation of LLMs on tasks requiring human-like cognition is a flourishing field of research (Sartori and Orrù 2023). A key area of focus is Theory of Mind (ToM), i.e., the ability to reason about others’ mental states (Kosinski 2024). Recent work has proposed a critical distinction between two forms of ToM: *literal* ToM, which is the ability to predict another agent’s behavior or beliefs; and *functional* ToM, which is the ability to use those predictions to rationally adapt one’s own behavior in an interactive context (Riemer et al. 2025). Our work is, to our knowledge, one of the first to apply this literal vs. functional distinction to the problem of belief revision in the context of human-aware AI. We demonstrate that while LLMs may show signs of literal competence by generating plausible explanations that resolve a contradiction, they may fail at using these explanations to guide a global, non-minimal revision of their human models.

User Studies on Belief Revision in Humans

To empirically investigate how humans resolve inconsistencies, we conducted three user study experiments designed to systematically probe the process of belief revision. Our methodology is grounded in established paradigms from cognitive psychology, which purposefully use structured, interpretable scenarios to isolate the cognitive mechanisms at play (Politzer and Carles 2001; Byrne and Walsh 2002; Khemlani and Johnson-Laird 2013). While the statements are intentionally simplified, this approach is a standard and necessary practice that allows for controlled experimentation with humans. To ensure our findings were not domain-specific, the scenarios were drawn from a variety of common, everyday contexts, including intuitive psychology, physics, and economics, and were designed to be highly plausible. The studies were conducted on the crowdsourcing platform Prolific (Palan and Schitter 2018), and participants were explicitly instructed that there were no right or wrong answers to encourage them to follow their natural thoughts.[2]

Our user studies examine how people reason when presented with information that contradicts their beliefs, resolving these inconsistencies into a new set of consistent beliefs. In particular, we conducted three experiments:

- Experiment 1 explores how people generate explanations when encountering inconsistencies, asking participants to explain why the new information conflicts with their existing beliefs;

[2] Ethics approval was obtained from our institution.

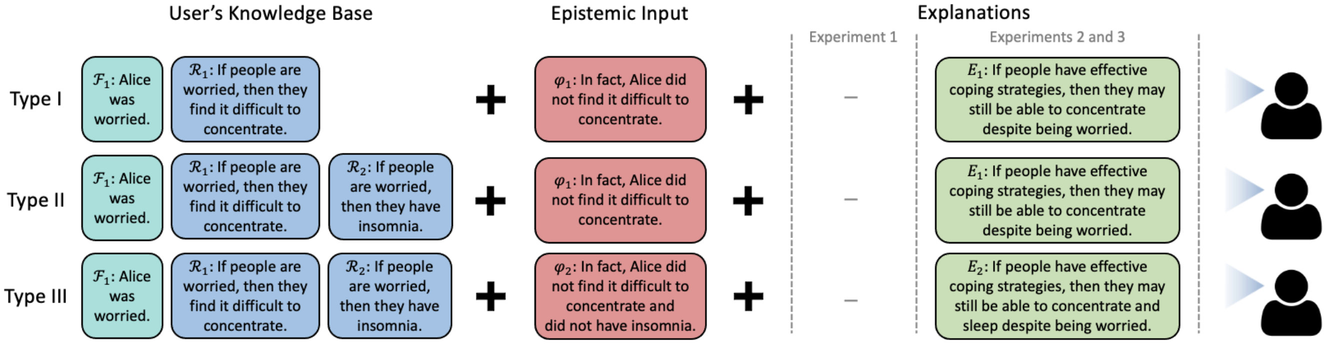

User's Knowledge Base
Epistemic Input
Explanations
Experiment 1
Experiments 2 and 3
Type I
$\mathcal{F}_1$: Alice was worried.
$\mathcal{R}_1$: If people are worried, then they find it difficult to concentrate.
φ_1: In fact, Alice did not find it difficult to concentrate.
–
E_1: If people have effective coping strategies, then they may still be able to concentrate despite being worried.
Type II
$\mathcal{F}_1$: Alice was worried.
$\mathcal{R}_1$: If people are worried, then they find it difficult to concentrate.
$\mathcal{R}_2$: If people are worried, then they have insomnia.
φ_1: In fact, Alice did not find it difficult to concentrate.
–
E_1: If people have effective coping strategies, then they may still be able to concentrate despite being worried.
Type III
$\mathcal{F}_1$: Alice was worried.
$\mathcal{R}_1$: If people are worried, then they find it difficult to concentrate.
$\mathcal{R}_2$: If people are worried, then they have insomnia.
φ_2: In fact, Alice did not find it difficult to concentrate and did not have insomnia.
–
E_2: If people have effective coping strategies, then they may still be able to concentrate and sleep despite being worried.

Figure 1: Experimental design for the three problem types, with examples of the human's knowledge base, comprising facts (turquoise boxes) and rules (blue boxes), and the epistemic input (red boxes), along with the explanations (green boxes) provided in Experiments 2 and 3.

- Experiment 2 examines how people revise their beliefs when provided with explanations (those generated in Experiment 1), testing whether explanation-based revision patterns persist when explanations are given rather than self-generated; and
- Experiment 3 investigates whether these revision patterns hold when beliefs are instantiated as specific cases rather than general rules, using grounded versions of the same scenarios with multiple concrete instances.

Design Overview: To carry out our investigation, we selected three types of problems, as shown in Figure 1. These problems present increasingly complex scenarios where new information (the epistemic input) conflicts with existing beliefs (the human's knowledge base), allowing a systematic study on how people handle such inconsistencies. More details for all experiments can be found in the supplement.

The first type (**Type I**) presents participants with a simple scenario containing a conditional (generalization) statement $\mathcal{R}_1$ and a (categorical) fact $\mathcal{F}_1$ about a specific case, which together compose the human's knowledge base, and the epistemic input φ_1 that conflicts with what the knowledge base implies. The second type (**Type II**) increases complexity by introducing an additional conditional statement $\mathcal{R}_2$, creating a scenario where the epistemic input φ_1 conflicts with the consequences of one of the conditional statements and the fact. Finally, the third type (**Type III**) of inconsistency presents the most complex case, where the more complex epistemic input φ_2 conflicts with the consequences of both conditional statements and the fact.

To analyze participant responses against the principles discussed in the introduction, we define two key concepts. First, we identify *explanation-based revision*, motivated by cognitive science research. Due to humans' propensity to envisage "disabling conditions", i.e., contextual factors that prevent a cause from producing its usual effect, their explanations are more likely to invoke such conditions than to imply that a categorical statement is wrong. For instance, an explanation for the conflict in Figure 1 is that "Alice has effective coping strategies." This explanation targets the universal applicability of the rule $\mathcal{R}_1$. We refer to this process as an explanation-based revision: a process where the revision is guided by a generated explanation, which naturally leads to a change broad modifications (such as modifying the general rule) to achieve explanatory understanding.

Second, to measure the outcome of this cognitive process, we define *non-minimality* in terms of informational loss, i.e., one that leads to larger informational loss than what is strictly necessary to restore consistency. For example, discarding the rule $\mathcal{R}_1$ implies rejecting all of its groundings, which means you can no longer infer that people find it difficult to concentrate if they are worried, for any instantiation of this rule. In contrast, discarding the categorical fact $\mathcal{F}_1$ leads to minimal information loss as it is a localized change that preserves the general rule.

Experiment 1

Participants and Design: We recruited 62 participants from Prolific across diverse demographics, with the only filter being that they are fluent in English. The participants carried out three different problems of each of the three types (Type I, Type II, and Type III), for a total of nine problems. The participants' main task was to explain the inconsistencies presented to them in their own words, specifically addressing why the new information (the epistemic input) could be true in light of the conflicting initial statements. After providing their explanations for every problem, each participant was asked how they approached explaining what was presented to them and if they followed any strategies when doing so.

Results: All participants easily came up with reasons to explain the inconsistencies they encountered. To analyze these free-text responses, we first examined their semantic content to infer which piece of the original information (e.g., categorical facts or conditional rules) the participant's explanation implicitly blamed for the conflict. Based on this inference, we employed a coding scheme that classified the implied revisions as either minimal or non-minimal based on our definitions in the previous section. For example, explanations implying non-minimal revisions included those that

	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%)	4.76×10^{-16}	1.28
	Type II	140 (86.96%)	21 (13.04%)	6.69×10^{-21}	1.48
	Type III	144 (81.36%)	33 (18.64%)	7.23×10^{-17}	1.25
	Aggregate	416 (83.37%)	83 (16.63%)	2.96×10^{-50}	1.33
Experiment 2	*Type I*	131 (79.39%)	34 (20.61%)	4.30×10^{-14}	1.25
	Type II	148 (91.36%)	14 (8.64%)	6.42×10^{-26}	2.27
	Type III	134 (85.90%)	22 (14.10%)	3.04×10^{-19}	1.68
	Aggregate	413 (85.51%)	70 (14.49%)	6.52×10^{-55}	1.64
Experiment 3	*Type I*	113 (71.97%)	44 (28.03%)	2.68×10^{-24}	2.10
	Type II	100 (65.36%)	53 (34.64%)	1.50×10^{-25}	1.79
	Type III	104 (64.20%)	58 (35.80%)	4.61×10^{-16}	1.52
	Aggregate	317 (67.16%)	155 (32.84%)	1.44×10^{-52}	1.45

Table 1: Results from all three experiments, with *Aggregate* representing combined data from all problem types.

introduced disabling conditions or directly negated a conditional rule (e.g., "It is not the case that if X then Y"). Explanations implying minimal revisions were those that negated the categorical fact (e.g., "Perhaps not X"). This scheme successfully classified 89% of all responses. The remaining responses either affirmed or denied the new information, i.e., the epistemic input, or were too vague to classify.

Table 1 displays the distribution of explanations implying either non-minimal or minimal revisions. The data reveal a compelling trend: a significant majority of explanations across all questions leaned towards revisions that imply removing or changing conditional rules. A Wilcoxon test performed on the aggregated data yielded a p-value significantly smaller than 0.05 ($p \approx 2.96 \times 10^{-50}$), providing robust evidence that the observed proportions are far from what would be expected by random chance. Moreover, effect size measurements (Cohen's d) were conducted to quantify the magnitude of these differences, where it was consistently high across all instances.

These results demonstrate that when faced with inconsistencies, participants predominantly created explanations that yield them to discard or modify conditional rules over categorical facts. This suggests that individuals engage deeply in resolving inconsistencies, often opting for more comprehensive explanatory frameworks that modify their existing beliefs to a greater extent than simply choosing an arbitrary minimal set of beliefs to remove.

Experiment 2

Participants and Design: In the second experiment, we recruited 60 new participants from Prolific with the same requirements as in Experiment 1. The goal of the study was to test whether the observed preference for non-minimal revisions persists when an explanation is provided to participants, rather than having them generate it themselves. Participants were presented with the same inconsistency problems, but this time they were also given one of the most plausible explanations (a disabling condition) generated by participants in Experiment 1. Participants were then asked to describe how they would revise their beliefs in light of this explanation.

Results: We employed a specific coding scheme to analyze how participants chose to revise their beliefs. In accordance with this scheme, participants indicated whether they would *keep*, *discard*, or *alter* the beliefs. When choosing to alter a belief, participants were asked to provide details about how they would go about it. As before, we have two categories of revision: non-minimal revisions that discard or alter either a conditional or a combination of more than two statements; and minimal revisions that discard or alter the categorical facts. This coding scheme classified 89% of the responses, while the remaining responses either yielded inconsistent revisions (e.g., not revising anything) or were too vague to be classified.

Table 1 and Figure 2 provide an overview of the results. The data reveal a clear trend: a significant majority of revisions were non-minimal across all problem types. Participants showed a strong preference for modifying or discarding conditional rules rather than categorical facts, with this pattern being particularly pronounced in Type II problems where over 90% of participants performed non-minimal revisions. The aggregate analysis across all 413 valid responses showed that 85.51% were non-minimal revisions. A Wilcoxon test performed on the aggregated data yielded a p-value of $p \approx 6.52 \times 10^-55$, providing strong statistical evidence that this preference was not due to chance.

These findings corroborate those of Experiment 1 and provide evidence that people predominantly opt for non-minimal revisions when presented with explanations. Even when given the explanations, participants maintained their tendency to make broader changes to their belief agents, suggesting that this preference for explanation-based revision may be a fundamental aspect of how people process inconsistencies.

Experiment 3

Participants and Design: Our final experiment recruited 60 new participants from Prolific (same requirements as before) and was designed to test the generality of our findings.

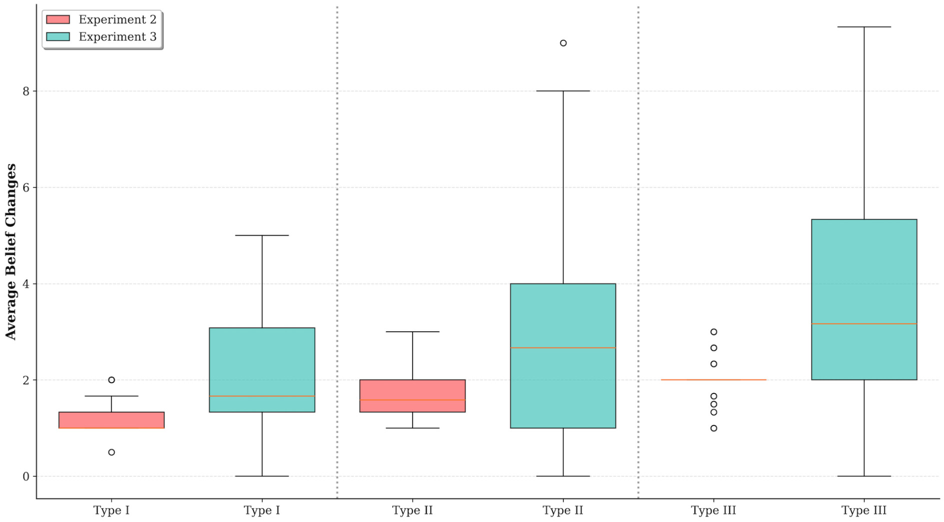


Figure 2: Distribution of average number of belief changes across problem types in Experiment 2 and Experiment 3.

We investigated whether the preference for non-minimal, explanation-based revisions would hold when the beliefs were not abstract statements but were instantiated as a set of specific, concrete cases. Particularly, for Type I scenarios (containing one generalization and one fact, we created four ground conditional statements, and added an additional categorical fact. For Types II and III scenarios (containing two generalizations and one fact), we created eight ground conditional statements (four for each rule), as well as added an additional fact for a total of two.

Results: Similarly to Experiment 2, we asked the participants to indicate whether they would keep, discard, or alter the beliefs in light of the explanation. However, given the instantiated nature of the beliefs, we established the following additional criteria for measuring minimal versus non-minimal revisions. For Type I and II scenarios, consistency could be restored with a single revision, i.e., either modifying one conditional rule or one fact. Therefore, any revision involving more than one belief change was considered non-minimal. For Type III scenarios, where the inconsistency affected two conditional statements, consistency required at most two revisions. Here, changes to more than two statements were considered non-minimal. To ensure consistent analysis, we counted both discarded and altered statements as changes in our measurement of belief revision. Using this coding scheme, we classified 87% of the responses, while the remaining responses yielded inconsistent revisions (e.g., keeping all beliefs).

As can be seen in Table 1 and Figure 2, our analysis revealed that participants consistently made more extensive changes than the minimum required for consistency. Type I scenarios showed a strong preference for non-minimal revisions, with 71.97% of participants opting for broader changes ($p \approx 2.68 \times 10^{-24}$). While these scenarios required only one revision for consistency, participants made an average of 2.21 changes to their beliefs, indicating a clear tendency to revise multiple statements rather than making minimal changes. The gap between minimally required to gain consistency and actual revisions became more pronounced in the subsequent scenarios. In Type II problems, which also required only one revision, 65.36% of revisions were non-minimal ($p < 10^{-25}$), with participants making an average of 3.08 changes. Type III scenarios, which required at most two revisions, showed 64.20% non-minimal revisions ($p < 10^{-16}$), with participants making an average of 3.94 changes—nearly twice the minimum required. This systematic increase in the average number of changes—from 2.21 in Type I to 3.94 in Type III—suggests that as scenarios become more complex, people make more revisions beyond what is minimally necessary to regain consistency.

These results provide more evidence that people's preference for non-minimal revisions persists even when dealing with concrete, grounded scenarios. These findings align with and extend the results from Experiments 1 and 2, demonstrating that whether working with general rules or specific instances, people consider and revise multiple related beliefs rather than making minimal, localized, or arbitrary changes.

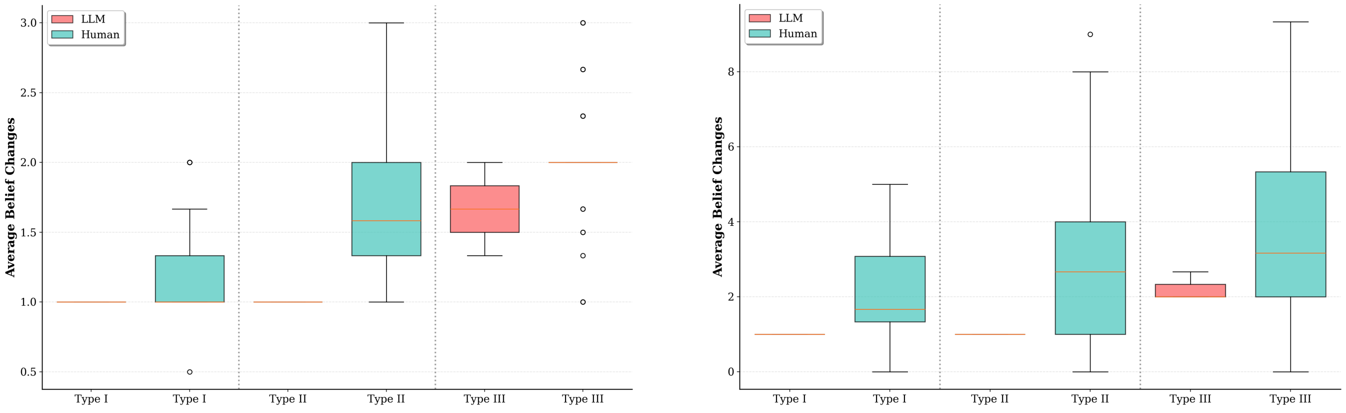
How Do LLMs Revise Human Models?

Having established a pattern of explanation-based revision in humans, we now examine whether today's advanced AI agents exhibit similar reasoning patterns when revising (given) models of human users. This investigation is aimed at understanding whether current LLMs, which are increasingly deployed in human-AI interaction settings, naturally align with humans' cognitive processes.

Recall from the introduction that, in the human-aware AI framework, an AI agent maintains both its world model $\mathcal{M}^R$ and a model of its human user $\mathcal{M}^H$, which, when faced with inconsistent information, has to be updated. Our central question is: do LLMs, acting as human-aware AI agents, perform this revision following an explanation-based revision process? This is important because, if there is a mismatch between how LLMs revise the human models and what hu-

Problem Type	**Task**	**Minimally Required**	**Human Avg. Changes**	**LLM Avg. Changes**		
				GPT o4-mini	**Claude 4 Opus**	**Gemini 2.5 Pro**
Type I	General	1	1.18	1.00	1.00	1.00
	Instantiated	1	2.21	1.00	1.00	1.00
Type II	General	1	1.63	1.00	1.00	1.00
	Instantiated	1	3.08	1.00	1.00	1.00
Type III	General	2	2.01	2.00	1.33	1.67
	Instantiated	2	3.94	2.00	2.00	2.67

Table 2: The average number of belief changes made by the human participants and each of the LLMs.



(a) General Task

(b) Instantiated Task

Figure 3: Distribution of average number of belief changes for (average) human and (average) LLM across problem types.

mans expect, then this could lead to model divergence (i.e., $\mathcal{M}_H$ progressively deviates from the human's actual beliefs), collaborative errors (i.e., incorrect predictions about human behavior), as well as hindering trust between the AI agent and the human user.

Experimental Setup: We selected three LLMs, OpenAI's GPT o4-mini, Anthropic's Claude 4 Opus, and Google's Gemini 2.5 Pro,[3] and presented them with the general scenarios (similar to Experiments 1 and 2) and instantiated scenarios (similar to Experiment 3) from the user studies. To answer our question, we evaluated if LLMs, when prompted to reason about a conflict, naturally generate their own explanations and revise the (given) human models in a manner that aligns with the explanation-based revision process observed in humans. Specifically, we followed recent work on LLM evaluation in cognitive tasks and employed a role-play framing by tasking the LLMs as human-aware AI agents needing to revise the human models in light of inconsistent information (Shanahan, McDonell, and Reynolds 2023).[4]

Note that to elicit a "thinking mode" in LLMs and log their reasoning process, we employed Chain-of-Thought prompting, explicitly asking the models to "think step-by-step" before providing a final revision decision, with deterministic output (temperature = 0) to ensure reproducibility. More details can be found in the supplement.

Results and Analysis: Table 2 summarizes the average number of belief changes made by humans and the LLMs, while Figure 3 visualizes the distribution of these changes.[5]

In the general task, the LLMs' behavior appears, at first glance, to align with human patterns. For Type I and Type II problems, both humans and LLMs averaged approximately one revision to resolve the conflict. An analysis of the models' reasoning reveals this single change was consistently a non-minimal one: they revised the general rule (R_1) rather than the fact (F_1), mirroring the human preferences we observed in our user studies. In the Type III general scenarios, which require two revisions to resolve the two-part contradiction, the results were more mixed. While GPT made the two required changes, Claude and Gemini made, on average, fewer (1.33 and 1.67, respectively), implying they sometimes failed to resolve the full inconsistency.

A different story emerges from the instantiated task. As illustrated in Figure 3(b), the LLMs' revisions are tightly clustered at the minimum required, while human revisions are

[3] At the time of writing this paper, these models are, arguably, the three of the best performing (reasoning) models.

[4] According to Shanahan et al., framing LLM behavior in terms of role play and simulation can allow us to draw on folk psychological terms observed in humans.

[5] To ensure robustness, each question instance was presented to each LLM 10 times, and the results were averaged.

broadly distributed and consistently higher. Humans make significantly more changes than minimally required, with the average number of revisions increasing with complexity from 2.21 to 3.94. In contrast, the LLMs followed a minimalist strategy. Interestingly, their reasoning traces frequently revealed this explicitly by stating that *"Humans make the smallest change necessary to resolve contradictions."* Moreover, in an instantiated scenario involving two entities (Alice and John), upon learning that "Alice did not lose concentration during her presentation," the LLMs would only revise the single rule pertaining to Alice and fail to generalize to the belief involving John.

Overall, the findings from this experiment suggest that while LLMs can simulate human-like explanation-based revision in simple contexts, their strategy might degrade to a minimalist approach as the scenario complexity increases. This indicates that their alignment with the human process of achieving explanatory understanding is not fully realized.

Discussion & Conclusion

Our empirical investigation reveals a mismatch between the principles of minimalism that have long guided AI and the psychological processes that govern human belief revision. Across three comprehensive user studies, we found robust and consistent evidence that people's revisions are driven by a search for explanatory understanding. This fundamental drive to understand why a conflict occurred leads individuals to make broad, non-minimal changes to their beliefs, systematically referring to modify general rules over isolated, specific facts. This pattern holds true whether people generate their own explanations (Experiment 1) or are provided with them (Experiment 2), and persists across both abstract, rule-based scenarios and more complex, instantiated ones (Experiment 3). This further highlights the fact that belief revision is not an isolated process, but an integral component of humans' broader quest for explanatory understanding.

Our evaluation of state-of-the-art LLMs reveals a potential misalignment with the cognitive process of explanation-based belief revision. While the LLMs demonstrated an ability to mimic explanation-based revision in simple, abstract scenarios, in the instantiated scenarios, they consistently reverted to making the minimum number of changes required to restore consistency.

Implications for AI Alignment: Our work may have implications for the safety and efficacy of human-aware AI agents, touching upon several core challenges in AI alignment. First, our findings expose a mistaken assumption in the human-aware AI paradigm. An AI agent that continuously updates its model of a human user using a minimalist operator is building its understanding on potentially faulty ground. As the human user makes explanation-based updates to their own beliefs, the AI agent's updates will fail to capture the scope and nature of these changes. This will lead to a compounding model divergence over time, where the agent's model of the human becomes progressively inaccurate, and unreliable.

Second, this divergence represents a vulnerability for scalable oversight (Ji et al. 2023). The goal of scalable oversight is to enable humans to supervise AI agents that may exceed their own capabilities, often by having AI agents learn from or interpret human feedback. Our results indicate that if the AI agent overseer operates on a flawed cognitive model of its human supervisor it may systematically misinterpret the very feedback it is designed to learn from. When a human provides feedback reflecting a broad, explanation-based change in their beliefs, a minimalist AI will likely attribute this to noise, irrationality, or a simple rejection of a single fact. It may misunderstand the intent behind the feedback, leading it to learn the wrong values or an incorrect model of the task, undermining the entire chain of supervision.

Third, our work shows the importance of "functional" cognitive alignment. An AI agent that can merely state a plausible explanation for a conflict (literal ToM) is not aligned if it fails to update its internal models and subsequent behavior in a manner consistent with that explanation (functional ToM). In general, as AI agents become more autonomous, their internal models of how humans think, decide, and react will become a critical determinant of their behavior. A misaligned cognitive model is a root cause of value misalignment. Before an AI agent can reliably learn what we want, it must possess an accurate model of how we think. Our research suggests that "Cognitive Model Alignment" must become a primary objective alongside "Value Alignment" in the broader research agenda.

Limitations & Future Work: We have to, of course, acknowledge the limitations of our study, which also point toward promising avenues for future research. Our experiments employed controlled, simplified scenarios, a standard and necessary methodology in cognitive psychology for isolating cognitive mechanisms. Future work should investigate whether these explanation-based patterns persist in more complex, naturalistic, and open-ended domains where beliefs are less structured and information is often ambiguous. Furthermore, our evaluation of LLMs represents a snapshot of a rapidly evolving technology; future model architectures may exhibit different patterns. The prompts used, though carefully designed, could also influence model behavior. Finally, it is crucial to remember that our experiments simulate belief revision in LLMs, which lack genuine, persistent mental states in the human sense.

A critical next step is to move from these empirical findings to a new class of formal, computational models. We plan to develop an explanation-based belief revision framework that provides a concrete, implementable operator for AI agents. Such a framework could be integrated into human-aware AI agents to enable more robust and accurate human modeling, connecting our work to research on logic-based formalisms for interactive XAI (Rago and Martinez 2024; Vasileiou et al. 2024). Moreover, we plan to explore methods for fine-tuning LLMs on datasets that explicitly demonstrate human cognitive patterns (such as those presented here). Finally, we plan to extend this work to dynamic, multi-turn interactions between humans and AI agents. This would allow us to study the compounding effects of model divergence over time and to test whether an AI agent equipped with an explanation-based revision operator can maintain better alignment with a human user.

References

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

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 156–163.

Elio, R.; and Pelletier, F. J. 1997. Belief change as propositional update. *Cognitive Science*, 21(4): 419–460.

Fermé, E.; Garapa, M.; Nayak, A.; and Reis, M. D. 2024. Relevance, recovery and recuperation: A prelude to ring withdrawal. *International Journal of Approximate Reasoning*, 166: 109108.

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

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; et al. 2023. AI alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Kambhampati, S. 2020. Challenges of Human-Aware AI Systems: AAAI Presidential Address. *AI Magazine*, 41(3): 3–17.

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

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

Makinson, D. 1997. On the Force of Some Apparent Counterexamples to Recovery. In Valdés, E. G.; Krawietz, W.; von Wright, G. H.; and Zimmerling, R., eds., *Normative Systems in Legal and Moral Theory*, 475–481.

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

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

Rago, A.; and Martinez, M. V. 2024. Advancing interactive explainable AI via belief change theory. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning (KR)*.

Riemer, M.; Ashktorab, Z.; Bouneffouf, D.; Das, P.; Liu, M.; Weisz, J. D.; and Campbell, M. 2025. Position: Theory of Mind Benchmarks are Broken for Large Language Models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

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

Sartori, G.; and Orrù, G. 2023. Language models and psychological sciences. *Frontiers in Psychology*, 14: 1279317.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role play with large language models. *Nature*, 623(7987): 493–498.

Son, T. C.; Nguyen, V.; Vasileiou, S. L.; and Yeoh, W. 2021. Model reconciliation in logic programs. In *Proceedings of the European Conference on Logics in Artificial Intelligence (JELIA)*, 393–406.

Sreedharan, S. 2023. Human-aware AI—A foundational framework for human–AI interaction. *AI Magazine*, 44(4): 460–466.

Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2021. Foundations of explanations as model reconciliation. *Artificial Intelligence*, 301: 103558.

Vasileiou, S. L.; Kumar, A.; Yeoh, W.; Son, T. C.; and Toni, F. 2024. Dialectical Reconciliation via Structured Argumentative Dialogues. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 777–787.

Vasileiou, S. L.; Previti, A.; and Yeoh, W. 2021. On Exploiting Hitting Sets for Model Reconciliation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 6514–6521.

Vasileiou, S. L.; Yeoh, W.; Son, T. C.; Kumar, A.; Cashmore, M.; and Magazzeni, D. 2022. A Logic-based Explanation Generation Framework for Classical and Hybrid Planning Problems. *Journal of Artificial Intelligence Research*, 73: 1473–1534.

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