# Proactive Counterfactual Inference in Flexible Decision Making

**Peiyue Liu**,**Weiwen Lu**, **Xiaohong Wan (xhwan@bnu.edu.cn)**
State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research
Beijing Normal University
Beijing, China

## Abstract

**In complex, uncertain environments, individuals must flexibly integrate multiple sources of information to adapt to changing task demands. While prior research has primarily focused on confidence formation and rule inference within a single task, less is known about how information across multiple tasks is integrated. Here,we designed an experiment to address this question by asking participants to infer task rules while switching between two tasks. We found that participants were able to maintain cognitive control in the face of task-irrelevant information, ensuring smooth task performance. However, when such irrelevant information could potentially support task rule inference, individuals can flexibly adjust their strategies, leveraging this information to optimize the decision-making process. Participants' beliefs about the current task rule (rule belief) modulated this cognitive flexibility, influencing how they prioritized, processed, and integrated information. Neural data revealed that the dorsal anterior cingulate cortex (dACC) plays a central role in these processes, specifically in: (1) encoding both task-relevant and task-irrelevant evidence; (2) updating rule beliefs and (3) modulating functional connectivity with the human fourth visual area and middle temporal area (hV4/MT). To probe the underlying mechanisms, we trained a recurrent neural network (RNN) model. We showed that within a trial, these neurons operate under an attention bottleneck, which serves as a constraint and mimics the potential attention-splitting process observed in humans. As with human participants, the effect of task-irrelevant information on rule belief updating was observed, but with a stronger effect. Together, these findings reveal a neural process in the human brain, particularly in the dACC, for integrating and updating beliefs about tasks, and how individuals flexibly adjust their strategies based on both relevant and irrelevant information.**

**Keywords:** cognitive flexibility; decision-making; counterfactual reasoning; confidence; anterior cingulate cortex; attention bottleneck

## Introduction

Humans are able to adaptively adjust strategies and allocate cognitive resources to optimize stimulus processing and meet the ever-changing demands of tasks. For example, when faced with stimuli containing multiple features, individuals typically focus their attention on task-relevant dimensions while suppressing or filtering out task-irrelevant information (Zanto & Gazzaley (2009)). However, studies on task belief in uncertain conditions and how such belief interacted with stimulus processing were lacking.

Previous research has primarily explored how task-relevant and task-irrelevant information interact within paradigms where task demands are clearly defined, thereby modulating cognitive processing (Flesch et al. (2022),Luo et al. (2025)). Studies using task instruction paradigms have revealed how the brain encodes both task-relevant and task-irrelevant information in a context-dependent manner (Moneta et al. (2023)), and how these two types of information are represented in brain regions such as the dorsal anterior cingulate cortex (dACC) (Ritz & Shenhav (2024b)) and dorsolateral prefrontal cortex (DLPFC) (Flesch et al. (2022)).

However, decision-making in the real world sometimes occurs in situations where task goals need to be inferred rather than explicitly stated. In such cases, the decision process involves two interconnected steps: inferring the current task state (i.e., task belief) and executing the corresponding task based on that belief. Task beliefs are updated through error-driven adjustments, influencing confidence in task-relevant information (Sarafyazd & Jazayeri (2019)), and dynamically optimizing rule beliefs (Xue et al. (2022)).Such hierarchical reasoning shapes the formation of rule beliefs, enabling the integration of perceptual information into task-relevant rules. Conversely, stronger rule beliefs enhance the discrimination of task-relevant perceptual features while leaving task-irrelevant features unaffected (Xue et al. (2022)). This interplay optimizes processing accuracy and promotes the efficient use of cognitive resources.

Notably, the role of task-irrelevant features in rule inference merits reconsideration. Traditionally regarded as sources of cognitive interference, these features may not be entirely irrelevant in rule-inference paradigms. They may correspond to alternative rules or provide valuable information for rule inference. How task-relevant and task-irrelevant information jointly contribute to rule-belief formation, and whether rule beliefs modulate the processing of task-irrelevant features, remains an unresolved research gap.

To investigate this, we designed a dual-task rule-inference experiment with measurement of both task confidence and rule belief using functional magnetic resonance imaging (fMRI) (Figure 1). Stimuli consisted of two rule-based features, with the chosen rule determining task-relevant features. We demonstrated that participants maintained robust cognitive control despite interference from task-irrelevant information, ensuring effective task execution. Notably, when task-irrelevant information could help rule inference, participants

utilized both task-relevant and task-irrelevant information to guide rule-based decisions, which was influenced by rule belief (Figure 2). The fMRI signals in the dACC exhibited sensitivity to (1) task-relevant evidence; (2) task-irrelevant evidence, and (3) rule beliefs. Decoding analyses further revealed that regions Human V4 (HV4) and Middle Temporal area (MT) encoded both task-relevant and task-irrelevant information during the task,modulated by rule beliefs. This collecting and utilizing of task-irrelevant information reflects proactive counterfactual reasoning, facilitating rule inference.

To test the computational plausibility of this mechanism, we developed a recurrent neural network (RNN) model with an attentional bottleneck as a normative model with minimal assumptions. During training, the model mirrored human behavior: under rule uncertainty, it proactively explored task-irrelevant features to seek evidence for alternative rules, supporting the hypothesis of proactive counterfactual reasoning as an optimal strategy (Figure 2).

## Method

### Task overview

Forty healthy adult participants (29 females, 11 males) performed a two-rule task during functional magnetic resonance imaging (fMRI) scanning (Figure 1). One participant was excluded due to misunderstanding the task button rules, resulting in a task accuracy below 50%. The final sample consisted of 39 participants. The experiment comprised two conditions: cue and no cue condition. Within each experimental block, the two rules (motion and color) had a 10% probability of switching per trial.

In the no-cue condition, participants were required to infer the task rule based on their judgment. Subsequently, they were asked to report their confidence in the predicted rule, with confidence levels ranging from 50% to 100%. In the cue condition, the task rule for each trial was explicitly instructed, displayed in a yellow box on the screen. Participants were required to confirm the cued rule by pressing a key. Following this confirmation, a randomly selected confidence level (50% to 100%) was presented as the default value, which participants were asked to confirm by pressing a key. This keypress ensured consistency in motor responses across cue and no-cue conditions, facilitating comparable fMRI measurements.

After the rule-confidence report, a random dot motion stimulus appeared in the center of the screen. Participants were asked to perform one of two tasks based on the inferred or cued rule: (1) judge the motion direction (left or right) of the random dot motion stimulus for the motion rule, or (2) judge the majority color (red or green) for the color rule.After the perceptual task decision, participants were also asked to report their confidence (ranging from 50% to 100%)in the decision. Subsequently, feedback was displayed on the screen, comparing the participant's perceptual decision with the correct rule and indicating whether the answer was correct.

For each trial, motion direction and color attributes were independently manipulated. Color consistency (i.e., the percentage of dots sharing the same color) and direction consistency (i.e., the percentage of dots moving in the same direction) were parametrically varied. Stimulus difficulty was calibrated during the training phase using a staircase procedure to achieve a target accuracy of 71%, which served as the intermediate difficulty level. Based on this, difficulty was then linearly adjusted to three levels, corresponding to approximate accuracy rates of 60%, 70%, and 80%. Consequently, participants in the fMRI experiment achieved an average stimulus accuracy of approximately 70%.

Each participant completed three sets of no-cue tasks and two sets of cue tasks, each consisting of 80 trials. The task order was counterbalanced across participants.

### RNNs With attention bottleneck

We trained recurrent neural networks (RNNs) in a supervised manner to perform a similar task. With unknown and switching rules, the RNNs was trained to predict the underlying rule and make task choice based on the chosen rule.With minimal assumptions, the RNNs were presented as normative models.

As in the experiment, there were two alternating rules switches with a hazard rate of 0.1 and there were inputs corresponding to different rules in each trial. To be detailed, each trial could be split into 3 phases, 3 steps of preparation phase, 7 steps of task phase, and 5 steps of feedback phase.

Throughout the trial, there would be a rule choice input indicating the chosen rule. During task phase, there would be stimulus inputs filtered by the attention bottleneck (see below). For each rule, there would be a correct answer with positive original strength. At the end of the task phase, RNNs needs to output a task choice to predict the correct answer based on the rule choice. During feedback phase, there would be feedback inputs based on the underlying rule and the task choice of RNNs as in the experiment. And at the end out the feedback phase, RNNs needs to output a rule choice which would be applied as the rule choice input of the next trial.

Compared to human participants, RNNs could process two different sources of information simultaneously unless there were further constraints. Here, we introduced an attention bottleneck to serve as a constraint and to mimic potential attention-splitting process in humans. The stimulus inputs corresponding to different rules was encoded in separate channels, while attention bottleneck was a weighting (summed to 1) on these channels so that channels for different rules had competing encoding strength. And after passed through the attention bottleneck, constant noise was applied on the stimulus inputs. This means that paying more attention to a certain stimulus feature may correspond to less attention to the other feature, and the feature with less attention would be more corrupted. In this setting, attention served as a bottleneck and the process of splitting attention actually reflected the "competing" processing of different features. For each rule, a channel corresponding to the ground truth answer would have an original stimulus strength randomly sampled from $U(0, 0.3)$, while the noises added at each step were sampled from $N(0, 0.1)$.

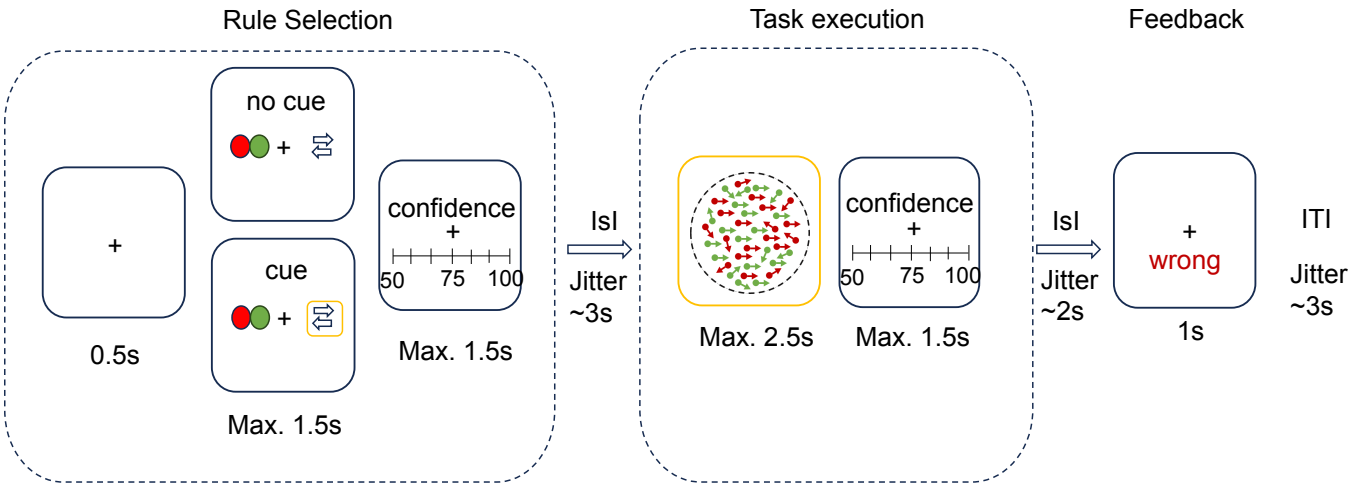


Figure 1: Task and design. Schematic of a dual-task experiment involving motion and color rule-based judgments under cue and no-cue conditions. **No-Cue Condition**: Participants predict the rule at the start of each trial and judge the stimulus accordingly. **Cue Condition**: Participants are given the rule at the start of each trial as a cue.

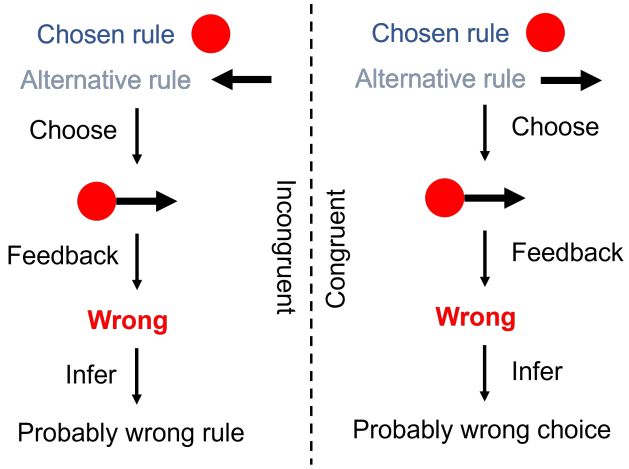


Figure 2: Schema for counterfactual reasoning in the task. If task-irrelevant information was collected, participants would be able to infer the validity of the alternative rule to help identify the ground-truth rule.

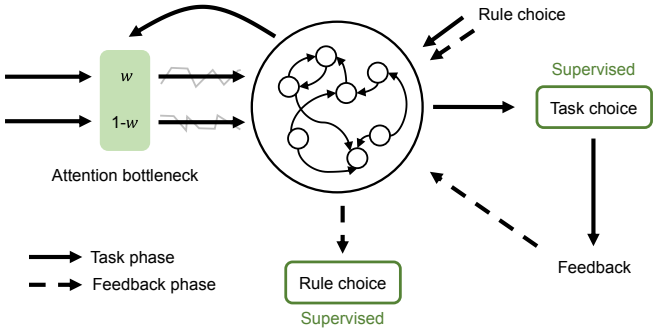


Figure 3: RNN framework.

40 RNNs were trained to match the number of subjects. The trained RNNs had 1024 hidden units, ReLU activation and linear readout for rule choice, task choice and attention, simple softmax was used to calculate corresponding confidence and attention weighting. We trained each RNN 2000 epochs with 100 trials in each epoch, with a batch size of 64.In other words, we simulated 64 sequences of trials in parallel, and each sequence had 100 consecutive trials. After each epoch, we used the ground truth rule and task answer for supervised learning, i.e. the loss function was loss $= l_{rule} + w_{task} l_{task}$, where $l_{rule}$, $l_{task}$ were cross entropy loss applied at the end of the feedback phase and task phase respectively and $w_{task} = 0.1$ was a weighting parameter. The weighting parameter was chosen heuristically so that trained RNNs could both solve the task and replicate the observed patterns in experiments (i.e., inferring the rule with the help of currently irrelevant information). While we admit the importance of the value of this parameter, it was not optimized as this was not the direct goal of this work. We used Adam optimizer with learning rate of 0.0001 implemented in PyTorch and gradient clipping with a maximum norm of 0.1 was also applied. When tested, 300 consecutive trials were simulated and coherence was discretized into 3 bins to match the 3 coherence levels in experiments post hoc.

## Results

### Task performance depends on relevant stimulus difficulty and rule confidence

The relevance and irrelevance of the stimuli were defined based on whether the stimulus attributes of the perceptual task were related to the current rule. For example, in trials where participants either judged the rule or received a color rule cue, stimuli related to color (such as red or green

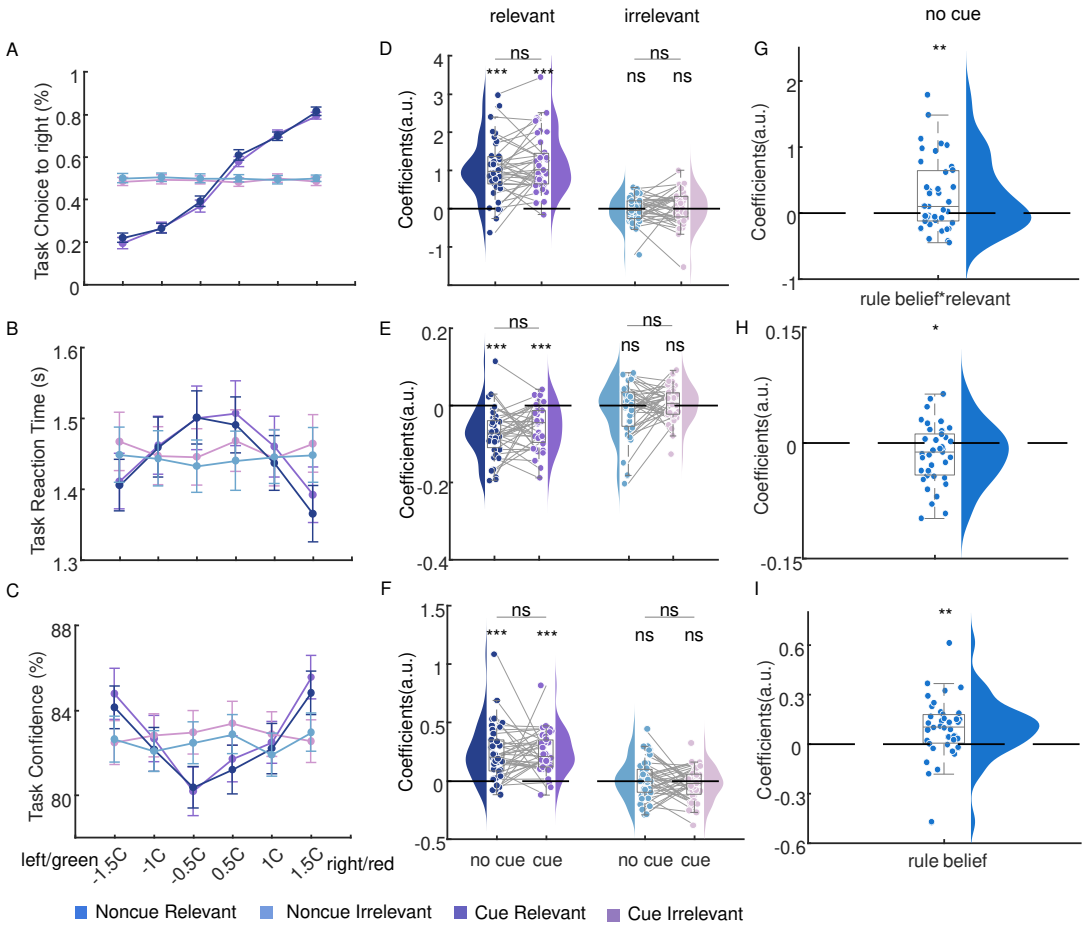


Figure 4: Task performance.The influence of relevant and irrelevant evidence on task choice preference (A), task reaction time (B), and task confidence (C).The regression effect of relevant and irrelevant evidence on task choice preference (D), task reaction time (E), and task confidence (F).The regression analysis weights of the interaction between rule confidence and relevant evidence on task choice preference (G) ; the weights of rule confidence in regression analysis on task reaction time(H) and task confidence(I).

major) were considered relevant, while stimuli related to motion direction (such as left or right motion) were considered irrelevant. We found that, under both the no-cue and cue conditions, participants' performance on the perceptual task was only sensitive to relevant information(Figure 4). When the consistency of relevant information was weak, participants exhibited lower choice accuracy (no-cue: $t_{(38)} = 8.25, p = 5.33 \times 10^{-10}$ cue: $t_{(38)} = 9.69, p = 8.08 \times 10^{-12}$), lower confidence (no-cue: $t_{(38)} = 6.27, p = 2.33 \times 10^{-7}$; cue: $t_{(38)} = 8.61, p = 1.81 \times 10^{-10}$), or slower reaction times (no-cue: $t_{(38)} = -7.26, p = 1.09 \times 10^{-8}$; cue: $t_{(38)} = -6.19, p = 3.16 \times 10^{-7}$). The results indicated no significant differences between the no-cue and cue conditions (paired-sample t-test: task accuracy: $t_{(38)} = -1.45, p = 0.15$; task confidence: $t_{(38)} = -0.08, p = 0.93$; task reaction time: $t_{(38)} = -1.31, p = 0.20$). Furthermore, the strength of evidence from irrelevant stimuli did not significantly affect participants' choice accuracy (no-cue: $t_{(38)} = -0.94, p = 0.36$; cue: $t_{(38)} = 0.14, p = 0.88$), task confidence (no-cue: $t_{(38)} = 0.35, p = 0.73$ cue: $t_{(38)} = -1.24, p = 0.22$), or reaction times (no-cue: $t_{(38)} = -1.06, p = 0.29$; cue: $t_{(38)} = 0.47, p = 0.48$). The analysis of the no-cue condition showed that participants' rule confidence had a significant impact on their task choices ( $t_{(38)} = 3.34, p = 1.9 \times 10^{-3}$)(Figure 4 G). As rule confidence increased, participants' task confidence also significantly improved ($t_{(38)} = 3.34, p = 1.9 \times 10^{-3}$)(Figure 4 I). Moreover, when participants had low rule confidence, their reaction times were slower ( $t_{(38)} = -2.38, p = 0.02$)(Figure 4 H).

## The impact of irrelevant evidence on rule belief update and switching

In no-cue conditions, we analyzed how relevant and irrelevant evidence influenced rule switching on the next trial using a regression model. We found that participants were more likely to stick with the current rule when more relevant evidence supported the correct answer ($t_{(38)} = -5.42, p = 3.55 \times 10^{-6}$), while they were more likely to switch rules with increased ir-

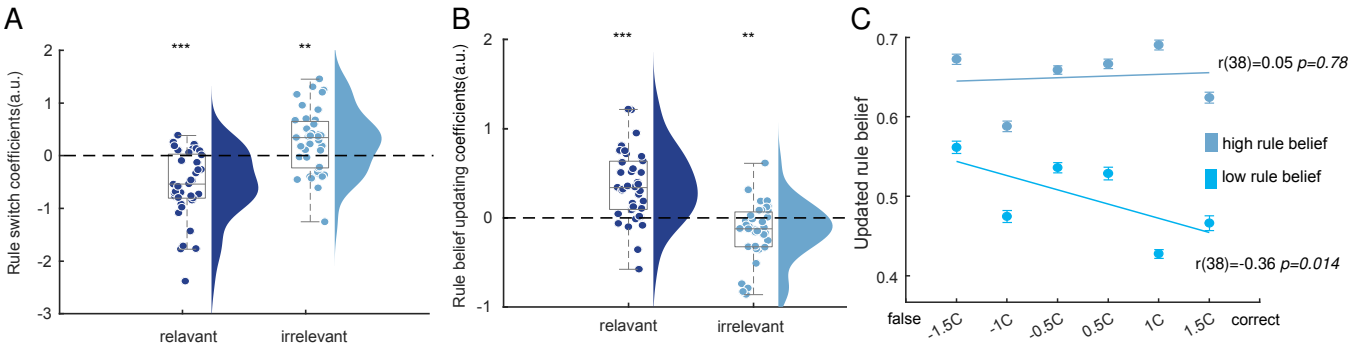


Figure 5: Rule performance.(A) Significant regression effect of relevant and irrelevant evidence on rule switch.(B) Significant regression effect of relevant and irrelevant evidence on rule belief updating.(C) Significant impact of irrelevant information on rule belief in the subsequent trial under low previous rule belief.

relevant information ($t_{(38)} = 2.77, p = 8.50 \times 10^{-3}$)(Figure 5 A). We observed that both relevant and irrelevant information significantly influenced the update of participants' rule confidence(Figure 5 B). Relevant evidence supporting the correct answer increased participants' rule confidence ( $t_{(38)} = 5.70, p = 1.49 \times 10^{-7}$), whereas irrelevant evidence supporting the correct answer shifted confidence toward the alternative rule ( $t_{(38)} = -2.94, p = 5.50 \times 10^{-3}$)(Figure 5).We further analyzed how irrelevant information affected rule confidence in the subsequent trial, depending on the previous trial's rule confidence. When participants had lower confidence in the current rule, irrelevant information supporting the correct option increased uncertainty and decreased rule confidence(Figure 5 C). In contrast, high rule confidence made updates less susceptible to irrelevant information ( $r_{(38)} = -0.36, p = 1.41 \times 10^{-2}$)(Figure 5). Specifically, when the previous trial's rule was correct, stronger irrelevant information led to a decrease in rule confidence, but only when rule confidence was low. This suggests that rule confidence plays a key role in how irrelevant information influences updates to rule confidence.

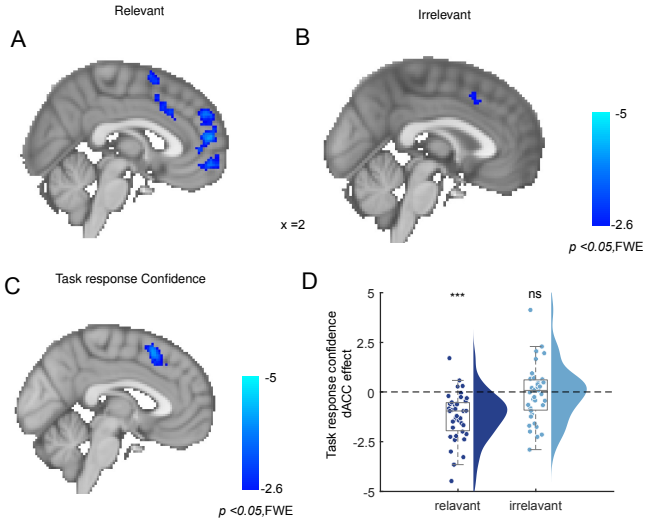


Figure 6: dACC representation in the Task Phase. (A)(B)(C) Representation of relevant evidence, irrelevant evidence and task response confidence, respectively. (D) Regression coefficients of relevant and irrelevant information on dACC

## dACC encoding of relevant and irrelevant evidence in decision-making

Our task design enabled us to investigate whether relevant and irrelevant information are encoded within the same region or whether they are processed separately during the task phases. Through a whole-brain generalized linear analysis, we observed that the dACC encodes both relevant and irrelevant information during task performance. The activation strength in the dACC showed a significant negative correlation with the strength of evidence from both relevant and irrelevant information (Figure 6). To investigate whether specific brain regions contribute to task-response confidence, we identified the dorsal anterior cingulate cortex (dACC) (MNI152: ±2, 21, 42) as a region of interest (ROI), based on its significant negative correlation with task-response confidence (Figure 6C). We extracted the neural activity time series from this ROI. Regression analysis revealed that task-relevant evidence was associated with a significantly negative beta coefficient for dACC activity related to rule-belief strength ( $t_{(38)} = -6.47, p = 1.30 \times 10^{-7}$), whereas task-irrelevant evidence showed no significant effect ( $t_{(38)} = 0.94, p = 0.94$). This finding suggests that the dACC plays a crucial role in decision-making by facilitating the integration of different types of evidence.

## dACC encodes rule confidence in task decisions, feedback, and rule decisions

Behaviorally, we observed that rule confidence influences task performance. We conducted a univariate analysis of rule-belief strength, examining whole-brain fMRI activity across the task decision, feedback, and rule decision phases. This analysis controlled for response confidence,reaction time, and feed-

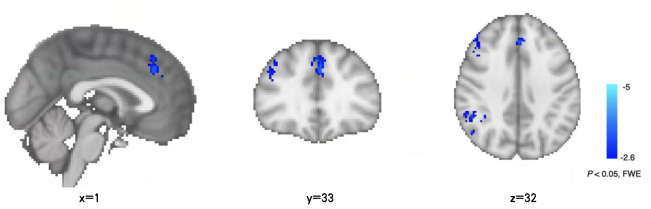


Figure 7: Conjunction of dACC's Representation of Rule Belief Across Phases

back. Additionally, we employed conjunction analyses to identify shared neural correlates of rule confidence across these three stages. The results revealed that fMRI activity in the dACC was negatively correlated with rule confidence ($z > 2.6$, $P < 0.05$, after cluster-level family-wise error (FWE) correction; (Figure 7)).

### Decoding irrelevant sensory information and the modulatory effect of rule confidence

To further investigate the neural representation of relevant and irrelevant evidence in the brain, we conducted a multivoxel pattern analysis (MVPA) based on neural activity sequences. The neural activity in the human V4 (hV4) and the middle temporal area (MT) was derived by computing the mean difference in neural activity between the color task and the motion direction task under cue conditions (i.e., color task minus direction task). These two regions were selected as the task-processing ROIs related to the color task (hV4: MNI152: (±22, -57, -16)) and the direction task (MT: MNI152: (±52, -60, 7)).We used 80% of the correctly performed trials under the cue condition as the training set, with the remaining 20% as the test set, and all trials under the no-cue condition as the test data. During training, we sorted the trials based on irrelevant key presses to ensure the training focused on irrelevant information, thus minimizing the influence of relevant information on the decoding process. We found that the neural activity in hV4 and MT could predict the relevant stimulus attributes with AUC above chance in both cue and no-cue conditions. However, only under the no-cue condition did neural activity in hV4 and MT significantly predict irrelevant stimulus attributes (Figure 8 A). Moreover, in MT, this prediction AUC was significantly higher than that under the cue condition ( $t_{(38)} = 2.55, p = 1.6 \times 10^{-2}$)(Figure 8 B). Additionally, decoding analyses of task stimuli in the human fourth visual area (hV4) and middle temporal area (MT) under no-cue conditions revealed distinct patterns modulated by rule-belief strength. In high rule-belief states, the area under the curve (AUC) for task-relevant feature discrimination was significantly higher than in low rule-belief states (MT:$t_{(38)} = 2.29, p = 2.8 \times 10^{-2}$,HV4:$t_{(38)} = 2.23, p = 3.3 \times 10^{-2}$)(Figure 8 E). Conversely, in low rule-belief states, the AUC for task-irrelevant feature discrimination in MT was significantly higher than in high rule-belief states ( MT:$t_{(38)} = -2.5, p = 1.7 \times 10^{-2}$)(Figure 8 E). These findings suggest that rule-belief strength modulates the neural processing of task-relevant and task-irrelevant features.

To examine whether rule confidence modulates dACC's processing of irrelevant information to MT and hV4, we conducted a psychophysiological interaction (PPI) analysis.We found that under high rule confidence, functional connectivity between dACC and hV4/MT was enhanced when processing relevant information. In contrast, under low rule confidence, functional connectivity between dACC and hV4/MT was enhanced when processing irrelevant information (paired t-test: hV4: $t_{(38)} = 4.39, p = 8.48 \times 10^{-5}$;MT: $t_{(38)} = 6.01, p = 5.01 \times 10^{-7}$). This suggests that, when rule confidence is low, dACC may be more involved in processing task-irrelevant stimuli (Figure 8 D).

### RNNs showed similar behavior patterns

Compared to human participants, RNNs had generally better performance and the task-irrelevant information also showed minimal effect on task performance while it could be significantly decoded (Figure 9 A,C). In fact, training the RNNs in a harder setting would lead to focusing on the relevant task which was different from the human pattern (See Figure S7 in supplementary material). It might reflect a different normative strategy.

A core pattern of human behavior was the effect of task-irrelevant information on rule belief updating. It corresponded to a normative explanation of counterfactual reasoning by testing the validity of the alternative rule. As the human participants, such effect was observed but with a stronger effect (Figure 9 D).

We were expecting a positive effect of rule belief on attention on task-relevant feature, as suggested by the human results. However, here the interaction term of previous rule belief and evidence showed a different pattern from human participants (See Figure S6 in supplementary material), which might be a result of generally high performance – where feedback might dominate rule belief updating, and attention had only slight effect on performance. However, following analysis of RNN attention pattern confirmed the proposed rule belief modulation in humans also existed in RNNs.

### Attention patterns revealed by RNNs

Despite no access to the actual attention pattern of human participants, the trained RNNs provided a chance to investigate the normative attention pattern in a similar setting.

Specifically, a positive effect of rule confidence on attention to relevant information was observed in the RNNs (12 RNNs with attention on relevant information $> 0.99$ excluded,$t_{(27)} = 2.90, p = 7.4 \times 10^{-3}$,Figure 10 right), which was consistent with the observed rule confidence modulation in human participants. This effect could reflect a proactive information seeking and counterfactual reasoning when the uncertainty is high.

Furthermore, RNNs tended to focus on the relevant task first and then switched to the irrelevant one (Figure 10 A). Additionally, the difficulty of the relevant task seemed to affect the

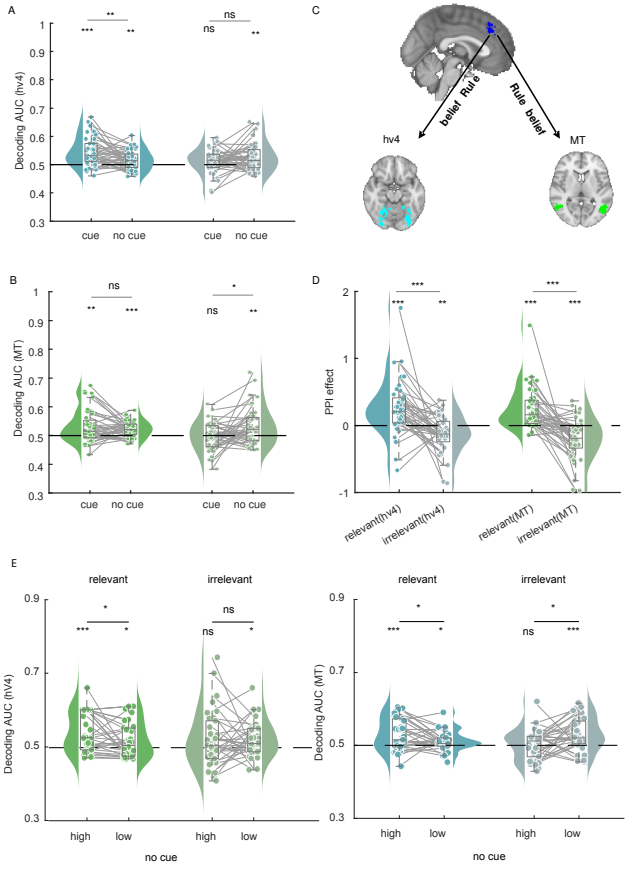


Figure 8: Decoding Task Stimuli in HV4 and MT, and dACC-Rule Belief PPI Analysis.(A) The human fourth visual area (hV4) significantly decodes color as task-relevant under both cue and no-cue conditions, and as task-irrelevant under no-cue conditions. (B) The middle temporal area (MT) significantly decodes motion direction as task-relevant under both cue and no-cue conditions, and as task-irrelevant under no-cue conditions. (C) Using dorsal anterior cingulate cortex (dACC) rule confidence as the seed region, the rule belief serves as the psychological measure for psycho-physiological interaction (PPI) analysis. (D) High rule belief significantly enhances hV4-dACC connectivity for the color rule, whereas low rule belief significantly enhances hV4-dACC connectivity for the motion rule; similar patterns are observed in MT. (E) Decoding of task stimuli in hV4 and MT under no-cue conditions, comparing high versus low rule-belief states.

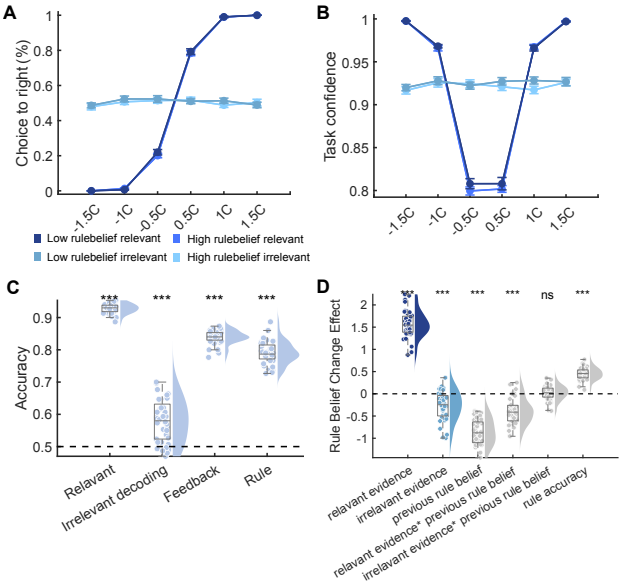


Figure 9: RNN behavior. (A) model task choice. (B) model task confidence. (C) model accuracy statistics. (D) model rule belief regression result.

attention-splitting, suggesting it is a dynamic process evolved together with decision-making (Figure 10 B).

Here, note that the RNNs were presented with fixed-duration stimuli and had no explicit reaction time, which was different from the human participants. Also, note that there were some trained RNNs only focusing on the relevant task (Figure 10 C), which might indicate diversity of strategies.

In summary, the RNNs demonstrated the proposed proactive counterfactual rule inference by splitting attention. Since the RNN results were generally consistent with human results, it suggested that humans also performed similar computation. And as the RNNs were proposed as a normative model, they showing attention-splitting behaviors modulated by rule belief suggested that these behaviors reflected an optimal strategy.

## Discussion

We examined how individuals perform rule-inference decisions in noisy, task-uncertain environments, uncovering both behavioral strategies and underlying neural mechanisms. We found that the participants were able to maintain sufficient cognitive control in the presence of irrelevant information interference to complete the task. This seemed to be different from former studies where distractor information interfere with current task, but a more difficult task setting and longer reaction time may explain for this Ritz & Shenhav (2024a). However, when such irrelevant information could positively influence task decisions, participants flexibly adjusted their strategies and used this information to optimize their decision-making process. Our decoding analysis further suggests that the hV4 and MT regions are involved in encoding rule-irrelevant information during task decisions, modulated by rule belief

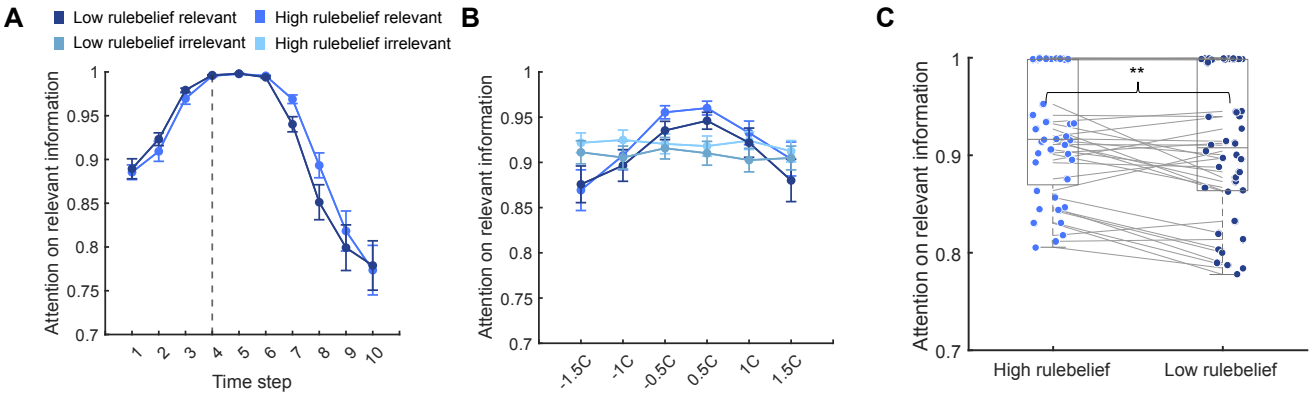


Figure 10: RNN attention patterns. (A) Average attention patterns within trials. (B) Average attention patterns with different stimulus coherence. With stronger task-relevant information, less attention on task-relevant information was applied. (C) High rule confidence corresponded to more attention on task-relevant information.

strength. The relationship between the representation of this irrelevant information and relevant information in the brain is reflected in the study by Moneta et al. (2023). They found that, in value-based decision making, vmPFC signals represent not only the relevant expected value, but also the irrelevant value from alternative contexts, with competition between the representations of relevant and irrelevant values. Furthermore, regarding the role of the dACC in the decision-making process, we observed that the dACC is not only involved in rule updating, but also plays a crucial role in regulating input connections. Specifically, high rule belief improves HV4-dACC connectivity for the color rule, while low rule belief enhances HV4-dACC connectivity for the motion rule, with a similar pattern observed in the MT region. We also found that the dACC representation of irrelevant information during task phases may be involved in the monitoring process.This suggests that the dACC supports individuals in making more flexible and adaptive decisions in the face of environmental uncertainty.

As the modulation of context signals on task feature processing has been generally observed and discussed (Moneta et al. (2023); Xue et al. (2022); Luo et al. (2025)), this attention-splitting explanation underscores the question of how context modulation relates to flexible decision-making. Typically, when rule inference was not required as in the cue condition, context modulation could be an inhibitory suppression of task-irrelevant information (Langdon & Engel (2025)). However, when rule inference was needed as in the no-cue condition and task-irrelevant information might help, different strategies and complex trade-off might exist. In such cases, the roles and interpretations of task interference, information seeking, cognitive control and all other related components need further investigation.

Using RNNs constrained by additional attention bottleneck, we replicated the core pattern of human behavior - task-irrelevant evidence did not interfere decision-making task while did help rule inference, suggesting a similar attention-splitting process in human flexible decision-making. Furthermore, RNNs were presented as a normative model with minimal assumptions. They naturally showed attention splitting behavior through training and demonstrated the proposed idea of proactive counterfactual inference as paying more attention to irrelevant feature to collect information when rule confidence was low. These results indicated that human behavior patterns and the modulation of rule belief on task features might originate from an optimal strategy.

In addition, the models predicted a pattern of normative attention. For example, they predicted evolving attention within trials and modulation of relevant task difficulty across trials. This reflected a process of real-time monitoring and attention-splitting, potentially linked to dACC. Hopefully, this idea could inspire advances in machine learning studies on the rule inference problem (Sommers et al. (2025)). Despite the complexity of decision-making strategies, future work could investigate the attention pattern with further experiments to test whether it is consistent with trained RNNs. However, it is worth noting that gaps between human behavior and RNNs still exist, which might indicate other unexplored characteristics of human cognition.

Altogether, our results suggested that participants may proactively collect and exploit the task-irrelevant information to support rule inference, probably through attention-splitting. These findings support the hypothesis of counterfactual reasoning, which enables individuals to selectively attend to relevant and irrelevant information depending on task demands. In dynamic environments, individuals exhibit high flexibility and adaptability, adjusting cognitive strategies to optimize decision outcomes in response to task complexity and uncertainty(Gilbert & Wilson (2007)).

## Acknowledgments

This research was supported by China National Science and Technology Innovation 2030-Major Projects-2021ZD0203700-1 (XW).

## References


Buckley, M. J., Mansouri, F. A., Hoda, H., Mahboubi, M., Browning, P. G., Kwok, S. C., ... Tanaka, K. (2009). Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. *Science*, *325*(5936), 52–58.

Dubreuil, A. (2022). The role of population structure in computations through neural dynamics. *Nature Neuroscience*, *25*. doi: 10.1038/s41593-022-01088-4

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543.

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258–1270.

Gilbert, D. T., & Wilson, T. D. (2007). Prospection: Experiencing the future. *Science*, *317*(5843), 1351–1354.

Hanks, T. D., & Summerfield, C. (2017). Perceptual Decision Making in Rodents, Monkeys, and Humans. *Neuron*, *93*(1), 15–31. doi: 10.1016/j.neuron.2016.12.003

Jahn, C. I., Markov, N. T., Morea, B., Daw, N. D., Ebitz, R. B., & Buschman, T. J. (2024). Learning attentional templates for value-based decision-making. *Cell*, *187*(6), 1476–1489.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231.

Langdon, C., & Engel, T. A. (2025, February). Latent circuit inference from heterogeneous neural responses during cognitive tasks. *Nature Neuroscience*, 1–11. doi: 10.1038/s41593-025-01869-7

Liu, Y., Xin, Y., & Xu, N.-l. (2021). A cortical circuit mechanism for structural knowledge-based flexible sensorimotor decision-making. *Neuron*, *109*(12), 2009–2024.

Luo, T., Xu, M., Zheng, Z., & Okazawa, G. (2025, January). Limitation of switching sensory information flow in flexible perceptual decision making. *Nature Communications*, *16*(1), 172. doi: 10.1038/s41467-024-55686-w

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, *503*(7474), 78–84.

Moneta, N., Garvert, M. M., Heekeren, H. R., & Schuck, N. W. (2023). Task state representations in vmpfc mediate relevant and irrelevant value signals and their behavioral influence. *Nature Communications*, *14*(1), 3156.

Pagan, M., Tang, V. D., Aoi, M. C., Pillow, J. W., Mante, V., Sussillo, D., & Brody, C. D. (2025). Individual variability of neural computations underlying flexible decisions. *Nature*, *639*(8054), 421–429.

Peterson, D. K., & Pitz, G. F. (1988). Confidence, uncertainty, and the use of information. *Journal of Experimental Psychology: Learning, memory, and cognition*, *14*(1), 85.

Qiu, L., Su, J., Ni, Y., Bai, Y., Zhang, X., Li, X., & Wan, X. (2018). The neural system of metacognition accompanying decision-making in the prefrontal cortex. *PLoS biology*, *16*(4), e2004037.

Ritz, H., & Shenhav, A. (2024a, March). Humans reconfigure target and distractor processing to address distinct task demands. *Psychological review*, *131*(2), 349–372. doi: 10.1037/rev0000442

Ritz, H., & Shenhav, A. (2024b). Orthogonal neural encoding of targets and distractors supports multivariate cognitive control. *Nature human behaviour*, *8*(5), 945–961.

Sarafyazd, M., & Jazayeri, M. (2019). Hierarchical reasoning by neural circuits in the frontal cortex. *Science*, *364*(6441), eaav8911.

Sommers, R. P., Thorat, S., Anthes, D., & Kietzmann, T. C. (2025). *Sparks of cognitive flexibility: self-guided context inference for flexible stimulus-response mapping by attentional routing.* Retrieved from `https://arxiv.org/abs/2502.15634`

Xue, C., Kramer, L. E., & Cohen, M. R. (2022). Dynamic task-belief is an integral part of decision-making. *Neuron*, *110*(15), 2503–2511.

Xue, C., Markman, S. K., Chen, R., Kramer, L. E., & Cohen, M. R. (2024, March). *Task interference as a neuronal basis for the cost of cognitive flexibility.* doi: 10.1101/2024.03.04.583375

Zanto, T. P., & Gazzaley, A. (2009). Neural suppression of irrelevant information underlies optimal working memory performance. *Journal of Neuroscience*, *29*(10), 3059–3066.

## Supplementary material

The previous RNN employed a shared neuron population for perceptual decision-making and rule inference. However, in human subjects, these functions correspond to separate brain regions. To achieve better structural alignment between the RNN model and human subjects, we developed a Decision-Inference RNN model (Figure S1). Compared to the earlier model, the Decision-Inference RNN model includes two distinct neuron populations: one for the decision module and another for the inference module. We retained the attention bottleneck to simulate human attention allocation constraints, while keeping the training tasks unchanged.Due to the separation of the perceptual decision-making and rule inference modules, the dependency between $l_{rule}$ and $l_{task}$ in the loss function decreases. Therefore, the loss function does not require weighted adjustment, and the training loss is simply $loss = l_{rule} + l_{task}$.

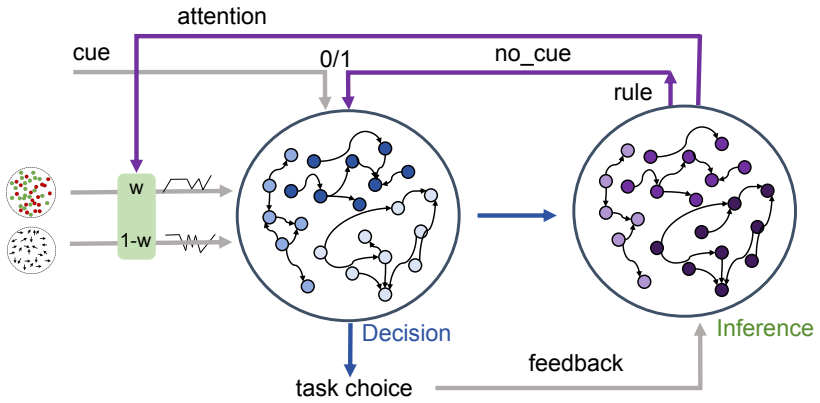


Figure S1: Decision-Inference RNN framework

Compared to human participants, RNNs generally exhibit better performance, and task-irrelevant information can be significantly decoded (Figure S2). We also observe the effect of counterfactual reasoning by testing the validity of the alternative rule (Figure S4). We calculated the average activation of each neuron at different time points under motion context and color context, sorted the neurons by the time of their peak activation, and plotted heatmaps for the decision module and inference module. The results show a clear separation of rule representations in the two modules(Figure S5).

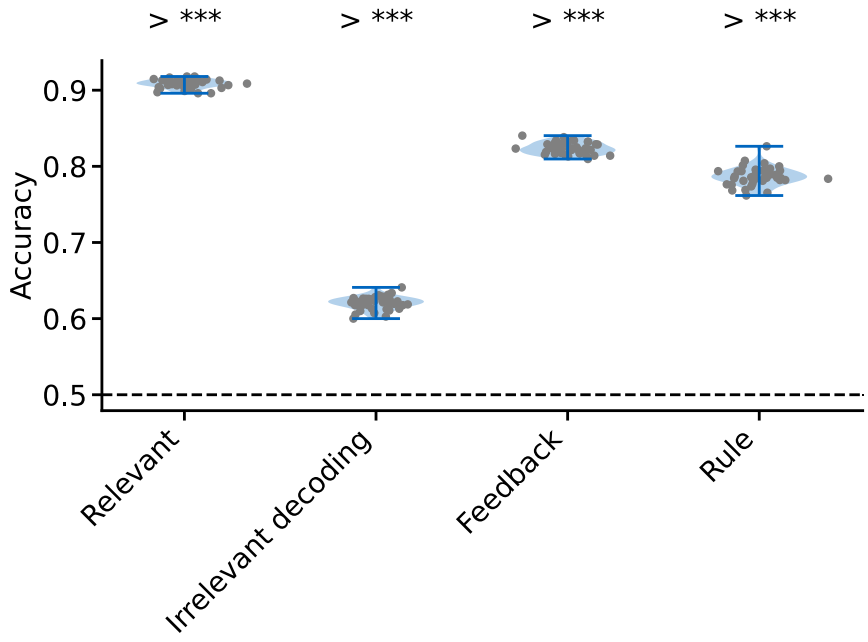


Figure S2: Model accuracy statistics

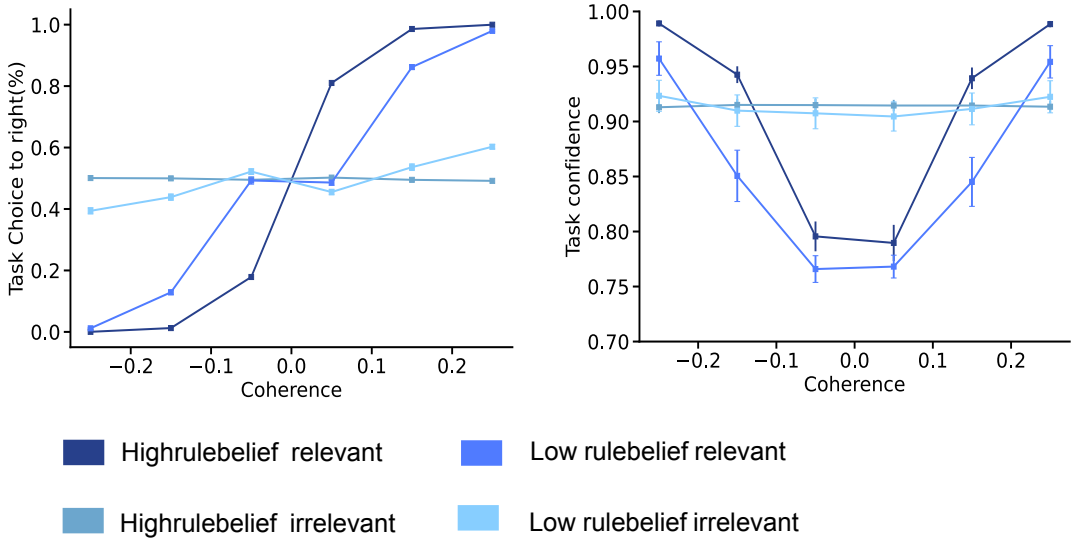


Figure S3: Model task choice and model task choice confidence

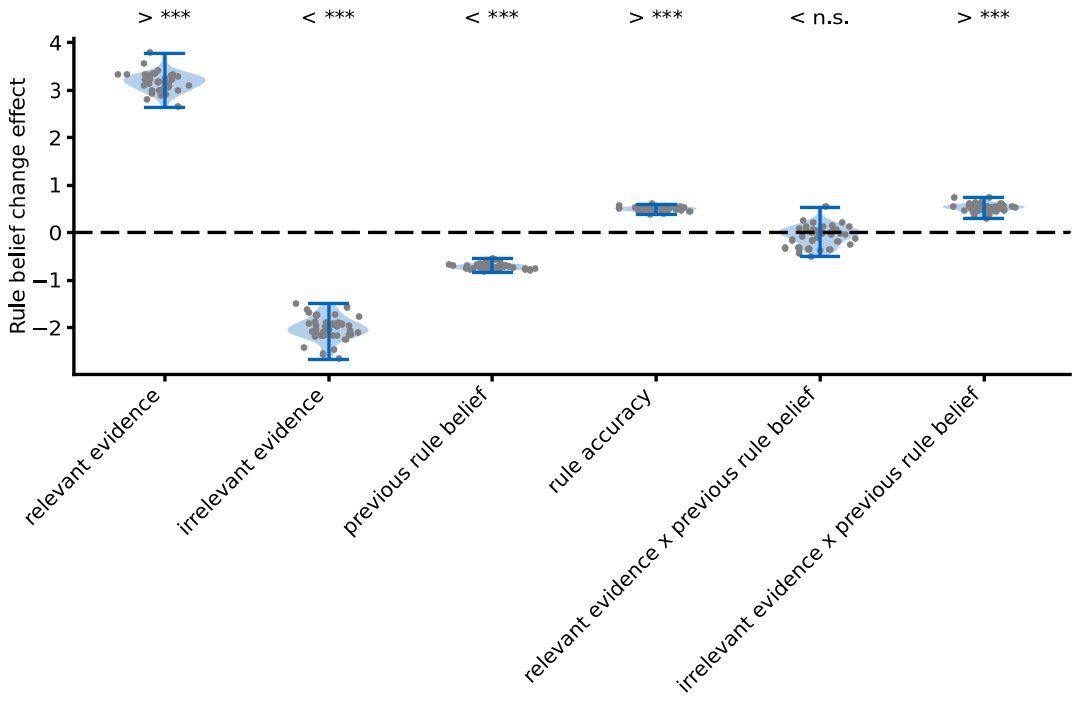


Figure S4: Model rule belief regression result

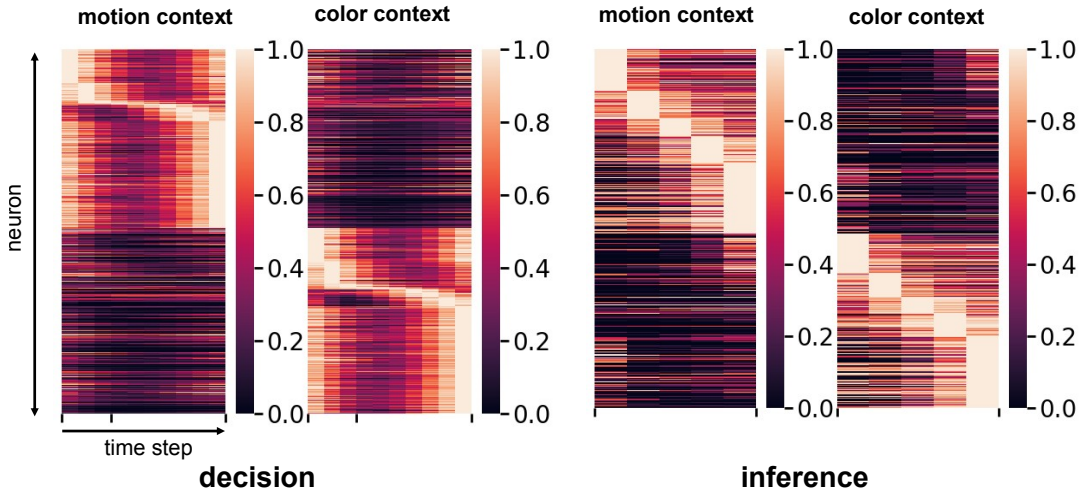


Figure S5: Neuron activations show a clear separation over time under the motion and color context

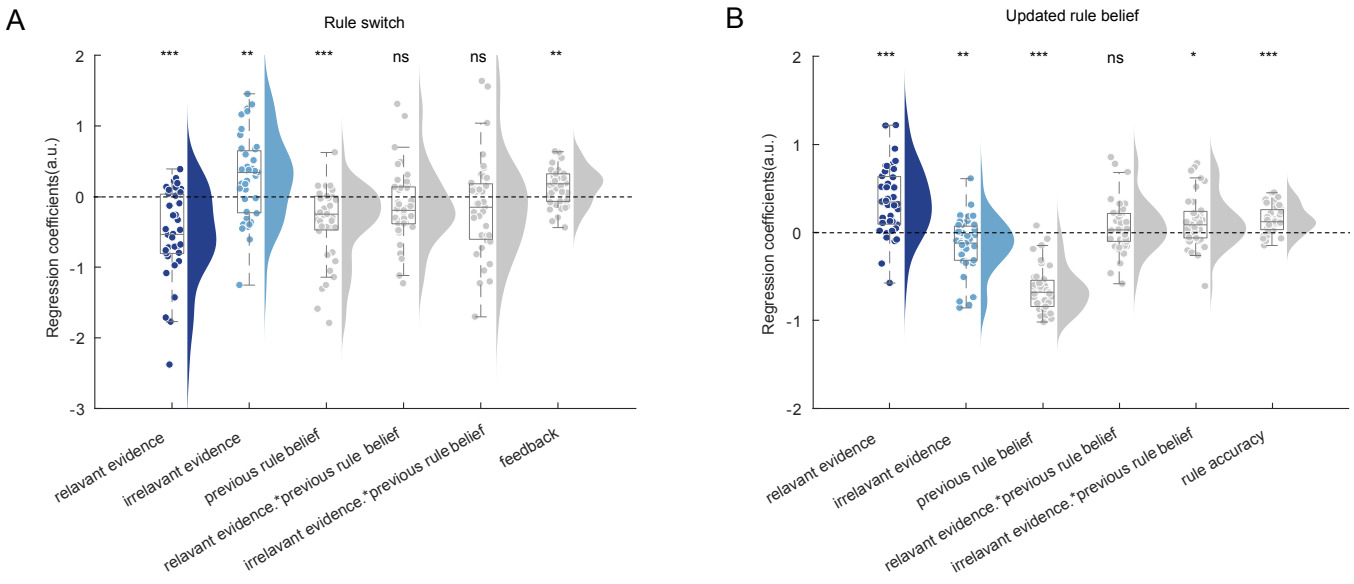


Figure S6: Rule switch and Updated rule belief regression analysis

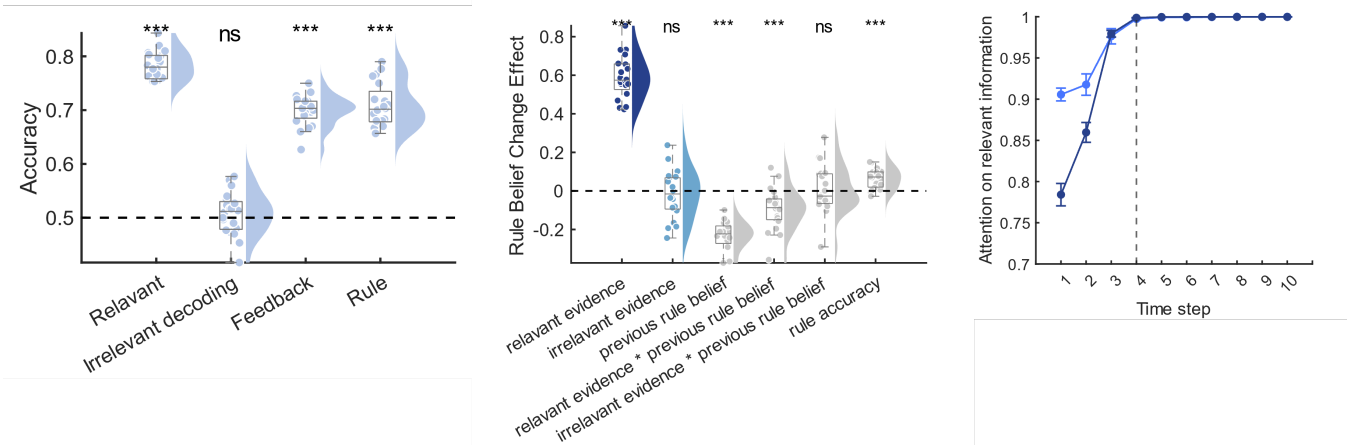


Figure S7: RNN results when trained in a harder setting (maximum coherence of 0.1). Despite better performance than human participants, trained models paid all attention to the relevant feature and no effect of irrelevant evidence could be seen. We think this might reflect a discrepancy of effort-accuracy relationship between human participants and RNN models. While RNN models could easily accumulate more evidence to make more accurate decisions in our setting, it might not be the case for humans.

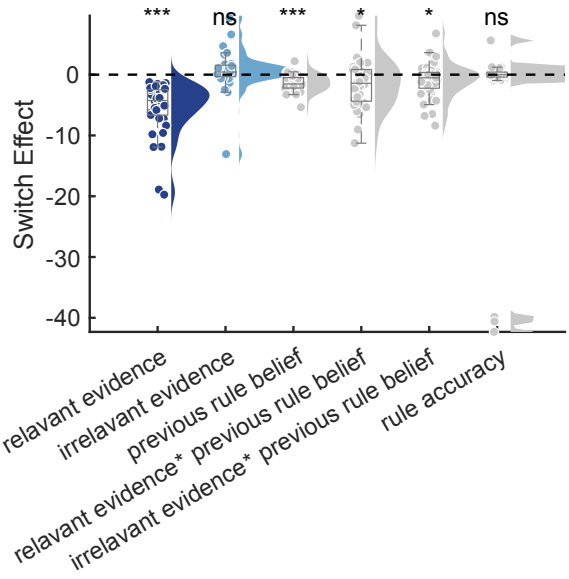


Figure S8: RNN switch GLM results. The analysis of rule switching of RNNs might be inappropriate as it was unbalanced (with a switch rate about 0.1). We think the rule belief could be a better metric in such analysis. Anyway,in such analysis the main effect of irrelevant evidence became insignificant (with an outlier), but the significant interaction term indicated the potential influence of irrelevant evidence.