Title

Adaptive Metacognition: The Interplay of Internal Beliefs and External Evidence in Dynamic Environments

ISBN

9798190944860

Author

LEE, RAIHYUNG

Publication Date

2026-09-11

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Adaptive Metacognition:

The Interplay of Internal Beliefs and External Evidence in Dynamic Environments

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Psychology

by

Raihyung Lee

2026

ABSTRACT OF THE DISSERTATION

Adaptive Metacognition:

The Interplay of Internal Beliefs

and External Evidence

in Dynamic Environments

by

Raihyung Lee

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2026

Professor Keith J. Holyoak, Chair

Adaptive cognition requires both persistence and flexibility: internal representations must remain stable enough to guide ongoing behavior, yet open enough to be revised when new evidence warrants it. This dissertation examines how this balance is achieved, advancing the view that the exchange between internal representations and external evidence is iterative rather than sequential. Internal states shape how incoming information is interpreted, and incoming information in turn prompts renewed processing — and revision — of the internal states themselves, in cycles that operate across timescales and levels of cognition and that are regulated by metacognitive signals of reliability. Three studies examine this iterative interplay in distinct settings. Study 1, using

human intracranial recordings, shows that perception of a brief input unfolds through a temporally extended comparison with recently formed memory, which repeatedly informs the developing perceptual decision rather than concluding in a single match. Study 2 shows that the iteration continues beyond commitment: feedback that conflicts with a confident judgment triggers a re-examination of the internal evidence supporting the original choice, allowing false feedback to be partly resisted. Study 3 shows that the exchange reaches the representations directing behavior: under uncertainty, people revise beliefs about their own goals in light of the observed outcomes of their own actions. Together, the studies indicate that internal representations and external evidence meet repeatedly rather than once, that revision proceeds by returning to what is held rather than overwriting it, and that metacognition — monitoring the reliability of each side and controlling whether another cycle of revision runs — is what keeps this balance adaptive in dynamic environments.

The dissertation of Raihyung Lee is approved.

Patricia Cheng

Hongjing Lu

Tao Gao

Keith J. Holyoak, Committee Chair

University of California, Los Angeles

2026

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my deepest gratitude to my advisor, Keith Holyoak, and to my former advisor, Hakwan Lau, for their generosity, support, and all that they have taught me throughout this journey. I am also sincerely grateful to my committee members, Patricia Cheng, Hongjing Lu, and Tao Gao, for their time, insight, and encouragement.

Finally, I want to thank all my dear friends who shared their lives with me along the way. I love you.

VITA

Raihyung Lee

Education

University of California, Los Angeles

M.A., Psychology, 2021

Seoul National University

M.S., Interdisciplinary Program in Neuroscience, 2018

B.A., Psychology, *summa cum laude*, 2016

Research Experience

Center for Neuroscience Imaging Research, Institute for Basic Science (IBS)

Post-Master's Researcher, 2025–2026

Center for Brain Science, RIKEN

Research Intern, August–December 2024

Teaching Experience

University of California, Los Angeles

Teaching Associate, 2022–2024

Teaching Assistant, 2019–2021

Seoul National University

Teaching Assistant, 2017–2018

Publication

Lee, R., Kwak, S., Lee, D., & Chey, J. (2022). Cognitive control training enhances the integration of intrinsic functional networks in adolescents. *Frontiers in Human Neuroscience, 16*, 8593

Chapter 1. General Introduction

Cognition requires a balance between information that is already represented internally and information currently provided by the environment. Perception is shaped by recent experience, decisions depend on accumulated beliefs and values, and behavior is guided by goals that persist beyond the immediate sensory moment. At the same time, these internal representations cannot remain insulated from new information. A belief that never changes in response to the world becomes inflexible, whereas a system that simply replaces its current state whenever new information arrives cannot make effective use of accumulated experience. Adaptive cognition therefore requires both persistence and flexibility: internal representations must remain sufficiently stable to guide ongoing processing while remaining sufficiently open to revision when new evidence warrants it — a cognitive expression of the stability–plasticity balance that adaptive systems of any kind must strike [1].

How this balance is achieved is often pictured, implicitly, as a sequence. Internal knowledge is applied to incoming information in a single pass and the result is deposited back into the store; a prior is combined with the current input, a commitment is made, and processing moves on [2–4]. A growing body of evidence suggests, however, that the exchange between internal and external is better described as iterative than sequential: not a single application of one term to the other, but a back-and-forth in which each side repeatedly re-engages the other, with influence running in both directions. In perception, the initial feedforward sweep supports only rudimentary analysis, and recognition of brief or degraded input is completed through recurrent exchange

between higher and lower areas [5, 6], while descending signals reshape early sensory responses even as those responses revise the expectations that generated them [7, 8]. Beyond perception, processing does not stop at commitment: evidence continues to accumulate after a choice, producing changes of mind without any new input [9, 10]. Consulting one's own representations a second time yields partially independent information, improving on the first reading [11, 12]; and what is processed after a decision is itself filtered through the decision that was made [13]. The internal term, in short, is not applied once and retired. It shapes how external evidence is read, and external evidence, in turn, prompts renewed consultation — and revision — of the internal term.

This iteration operates at multiple timescales and at multiple levels of cognition. At the briefest timescale, the loop closes within a single act of perceiving or deciding, over the course of hundreds of milliseconds, as a stimulus representation and an internal expectation are brought into register through repeated cycles rather than at a stroke. At longer timescales, the products of one exchange become the starting points of the next — today's posterior is tomorrow's prior — and in changing environments the pace of revision is itself adjusted: learners update more rapidly when contingencies are volatile and more conservatively when they are stable [14–17]. The brain appears organized for exactly this nesting, with intrinsic processing timescales that lengthen along the cortical hierarchy, embedding fast sensory loops within slower circuits that maintain context and belief [18, 19]. Nor is the iteration confined to the interpretation of sensory input. The same back-and-forth appears at higher levels of cognition, where the internal term is not an expectation about a stimulus but a belief the agent holds — about what was decided, about what is worth pursuing. When the world pushes back against such a belief, revision need not be a simple overwrite of the current state: it can involve returning to the evidence and experience

from which the belief was built, so that the exchange with the environment reaches backward into the representations themselves.

A loop, unlike a line, needs governance. If the exchange between internal and external can always run another cycle, something must determine whether a further cycle is warranted, how far the current state should move, and which side of the exchange deserves more trust at a given moment. This regulatory problem is metacognitive in nature, and decision confidence — the best-studied metacognitive variable — has exactly the profile the job requires: it is well described as an estimate of the probability that one's judgment is correct [20, 21], expressing the standing of an internal state in a currency directly comparable with external evidence. And confidence demonstrably exercises this control: it determines whether further information is sought [22], how strongly new evidence bears on learning [23], and how much outside input is absorbed into one's own judgment [24]. The second-order framework of metacognition [25] explains how such regulation is possible at all: metacognitive judgment is itself an inference over one's own states, drawn from imperfect internal evidence — which is why it can genuinely inform the exchange, and also why it can err. In this thesis, adaptive metacognition names this regulatory function: the monitoring of how much the current internal state and the current external evidence each deserve, and the control of whether — and how far — another cycle of revision runs. It is the mechanism by which persistence and flexibility are balanced, moment to moment, in environments that do not stand still.

The three empirical studies of this dissertation examine this iterative interplay in three distinct settings. They differ in domain, in method, and in the level of cognition they address; they are offered as three windows onto the same broad phenomenon rather than as a single experiment in

three forms. Study 1 (Chapter 2) observes the iteration directly, at its fastest timescale, in perception. Participants judged a briefly presented image against the memory of a naturalistic episode formed moments earlier, while intracranial recordings tracked the underlying computations. The comparison between the internally-maintained representation and the current input was not completed in a single step: information about their correspondence emerged early, persisted, and repeatedly informed the developing perceptual decision over the course of the trial.

Study 2 (Chapter 3) examines the iteration as it runs beyond a committed decision. Participants made a difficult perceptual choice, rated their confidence, received feedback of covertly varying reliability, and rated their confidence again. Their revisions went beyond any single-pass weighting of the feedback: when feedback conflicted with a confident judgment, participants re-examined the internal evidence behind their initial choice, and this renewed consultation allowed them to partly resist feedback that was in fact false — with confidence governing whether the additional cycle ran at all.

Study 3 (Chapter 4) follows the loop to a higher level of cognition, where external evidence feeds back into the internal states that direct behavior. In a value-guided task, participants acted to reach reward-bearing options and then reported which option they had wanted. What they observed of their own action's outcome was incorporated into that report — most strongly when volatile reward contingencies had weakened their value-based expectations — a pattern captured by a Bayesian model in which the observed outcome serves as evidence about one's own goal. Here the exchange completes the circle: information from the environment revises the internal representation from which subsequent behavior proceeds.

Together, the studies motivate a modest integration rather than a unified mechanism: across perception, decision evaluation, and value-guided action, internal representations and external evidence meet not once but repeatedly, at timescales from milliseconds to the run of learning, and metacognitive signals help decide how each round of the exchange is settled. Chapters 2–4 present the studies in full; Chapter 5 returns to these themes, tracing what links the three studies and what remains particular to each.

References

1. Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review, 87*, 1–51.
2. Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*, 712–719.
3. Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30*, 535–574.
4. Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*, 873–922.
5. Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences, 23*, 571–579.
6. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience, 22*, 974–983.
7. Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*, 79–87.
8. Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience, 14*, 350–363.
9. Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature, 461*, 263–266.
10. Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*, 864–901.
11. Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*, 645–647.

12. Elosegi, P., Rahnev, D., & Soto, D. (2024). Think twice: Re-assessing confidence improves visual metacognition. *Attention, Perception, & Psychophysics, 86*, 373–380.
13. Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., & Donner, T. H. (2018). Confirmation bias through selective overweighting of choice-consistent evidence. *Current Biology, 28*, 3128–3135.
14. Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience, 10*, 1214–1221.
15. Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience, 30*, 12366–12378.
16. Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience, 5*, 39.
17. Soltani, A., & Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience, 20*, 635–644.
18. Murray, J. D., Bernacchia, A., Freedman, D. J., Romo, R., Wallis, J. D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D., & Wang, X.-J. (2014). A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience, 17*, 1661–1663.
19. Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology, 4*, e1000209.
20. Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience, 19*, 366–374.
21. Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron, 88*, 78–92.
22. Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science, 29*, 761–778.

23. Meyniel, F., & Dehaene, S. (2017). Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences, 114*, E3859–E3868.
24. Pescetelli, N., & Yeung, N. (2021). The role of decision confidence in advice-taking and trust formation. *Journal of Experimental Psychology: General, 150*, 507–526.
25. Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124*, 91–114.

Chapter 2. Neural Dynamics of Memory–Sensory Comparison in Perceptual Detection

1. Introduction

Perception is not determined solely by information currently available to the senses [1, 2]. Because sensory signals are often brief, noisy, or incomplete, the brain often interprets them in relation to information retained from the recent past [3–5]. A key way in which recent information influences perception is through working memory, which keeps previously encountered information accessible during the processing of subsequent sensory input [6, 7]. Consistent with this role, behavioral studies have shown that information maintained in working memory can facilitate the processing of corresponding visual stimuli and alter subsequent perceptual judgments [8–11]. Complementing these behavioral findings, neural studies indicate that maintained visual information can remain represented within perceptual systems after the original sensory input has disappeared and can modulate responses to subsequently presented matching stimuli [12–15]. Together, these findings support an interactive account in which working memory does not merely preserve the output of perception but actively shapes the processing of subsequent sensory information.

The computational process underlying this interaction, however, remains incompletely understood. One plausible account is that a newly presented stimulus is compared over time with an internally maintained representation in working memory, with information about their

correspondence progressively contributing to stimulus identification. Several findings support the idea that the brain evaluates incoming sensory input in relation to working-memory contents and that correspondence between the two facilitates perceptual processing. Differences between representations held in visual working memory and newly presented perceptual inputs can be detected efficiently [16]. Moreover, stimuli corresponding to working-memory contents are more likely to reach visual awareness and can be identified with greater perceptual sensitivity than noncorresponding stimuli [9, 17, 18]. Model-based evidence further suggests that matching stimuli require less evidence to reach conscious perception [10], while neural responses to such stimuli are enhanced both in overall magnitude and in the strength of stimulus-specific representations [15]. Together, these studies demonstrate that correspondence between internal and external representations shapes ongoing perceptual processing. They do not, however, reveal the spatiotemporal dynamics through which this correspondence is evaluated—that is, when and where comparison-related information emerges and how it develops across time and brain regions to support stimulus identification.

Research on perceptual decision-making has established that perceptual judgments often emerge not in a single step, but through the progressive integration of noisy, task-relevant information over time [19–21]. Consistent with this view, neural signals related to perceptual decisions vary with sensory evidence and develop gradually toward perceptual recognition and detection [22–24]. Extending this logic, if working memory contributes to perception by relating current sensory input to internally maintained information, this comparison may likewise unfold over time rather than occur instantaneously, continuously supplying information that contributes to perceptual identification. Such temporally extended processing may be particularly important when the external stimulus is brief, because the processes supporting identification must

continue after the sensory input has disappeared [25, 26]. Under these conditions, the brain may continue to process a sensory representation of the stimulus, compare it over time with information maintained in working memory, and use the ongoing relationship between the two to support identification. We therefore hypothesized that comparison between internal and external representations would emerge early and persist over time, continuing to provide information that supports the formation of a perceptual judgment.

The integration of internal and external information may depend on coordination among large-scale cortical systems. Recent work has implicated the default mode network (DMN) in constructing and using representations that extend beyond information immediately available to the senses. DMN regions integrate information across relatively long temporal windows [27, 28], enabling construction of coherent representations of events or situations that extend across perception and memory [29, 30]. By maintaining such representations beyond the immediate sensory moment, the DMN allows previously encountered or internally generated information to guide ongoing cognition [31]. These properties may allow the DMN to maintain broader contextual representations that shape ongoing perception and decision-making. Such representations may provide the internal context against which newly arriving sensory information is evaluated [32].

Memory-guided cognition is unlikely, however, to be implemented by the default network alone. Tasks requiring decisions based on internal information can jointly recruit default-network and frontoparietal regions, suggesting that internally maintained representations must be coordinated with systems supporting task-relevant control and selection [31]. In perceptual decision-making, decision-related information emerges across widely distributed cortical regions rather than within

a single dedicated system [23], with frontoparietal regions among those most consistently implicated in representing and integrating information relevant to an impending choice [20, 21, 33]. One possibility, therefore, is that regions associated with the default network contribute more strongly to representing or evaluating the relationship between current input and internally maintained context, whereas information produced by this comparison becomes increasingly distributed across cortical systems involved in perceptual decision-making, with particularly strong involvement of frontoparietal regions.

To examine how these computations unfold across time and brain regions, we used stereo-electroencephalography (sEEG), which provides millisecond-scale temporal resolution from anatomically localized recording sites. Participants viewed animal videos lasting approximately 4.5 s, followed by a delay of approximately 1.5 s and a 100-ms picture drawn either from the preceding video or from a different video. They then judged whether the picture contained a primate or a non-primate animal. This design allowed us to test whether correspondence between a recently formed video representation and a briefly presented sensory input facilitated perceptual identification, and to track when and where information about memory–sensory correspondence emerged and how it related over time to neural representations of the picture's category.

To anticipate the findings, participants categorized pictures more accurately when they were drawn from the preceding video than when they were drawn from a different video. Neural analyses suggested that representations of the briefly presented picture were continuously evaluated in relation to a representation of the preceding video and that comparison-related information contributed to the subsequent development of information about the picture's

category. Comparison-related signals were most prominent in electrodes located in regions associated with the DMN, whereas the later development of evidence supporting perceptual identification was distributed broadly across cortical regions, with relatively greater frontoparietal involvement. Together, these findings reveal a temporally evolving neural process through which internally maintained information is related to a transient sensory representation and progressively contributes to perceptual identification.

2. Methods

2.1. Participants

Intracranial recordings were obtained from 19 patients with intractable epilepsy who underwent implantation of intracranial electrodes for clinical evaluation and treatment at the Second Affiliated Hospital of Zhejiang University School of Medicine (SAHZU). All participants were right-handed and had normal or corrected-to-normal vision. Eight participants were excluded because their data indicated that they had misunderstood the task instructions, and one additional participant was excluded because of poor sEEG signal quality. The final sample therefore consisted of 10 participants (mean age = 23.4 years, SD = 7.78; 3 female). The study was approved by the institutional review boards of Duke Kunshan University (FWA00021580) and SAHZU (IR2020001335). No clinical seizures were observed during the experiment.

2.2. Experimental design

Participants performed a perceptual detection task following the presentation of a short video, designed to test whether recently viewed visual information facilitated identification of a briefly presented picture (Figure 2-1a). Trial initiation was self-paced: participants began each trial by pressing a start button. Each trial started with a trial-unique video clip depicting animals from either the primate or non-primate category and lasting 4–5 s. After the video ended, a blank delay period of approximately 1.5 s followed. A picture was then presented for approximately 100 ms, and participants were instructed to indicate as quickly as possible whether it contained a primate by pressing the corresponding response button. Half of the pictures contained a primate, and the other half contained a non-primate animal. On one-fifth of all trials, the picture was drawn from the video presented on that trial; on the remaining four-fifths, it was drawn from a different video. Participants were not informed that some of the pictures would be drawn from the preceding video. Participants completed a mean of 168.5 trials, with the total number ranging from 87 to 213 across participants. Trials were divided into multiple sessions of up to 40 trials each. Breaks were provided between sessions, during which participants were instructed to rest. Following each perceptual decision, participants completed a temporal-order judgment task and provided a confidence rating. These measures were not analyzed in the present study and will be reported separately. Before the main experiment, all participants completed a training session using the same experimental program but different stimulus materials.

2.3. sEEG Data Acquisition and Signal Processing

2.3.1. Intracranial recordings and electrode localization

Stereo-electroencephalography (sEEG) data were recorded using stereotactically implanted depth electrodes. Electrode contacts consisted of 2-mm platinum cylinders with a diameter of 0.8 mm and an intercontact spacing of 3.5 mm (Sinovation, Beijing, China). All electrodes were implanted solely for clinical diagnostic purposes as part of the patients' evaluation for neurosurgical treatment of epilepsy. Participants were implanted with a mean of 11.8 depth-electrode shafts each (range = 8–14), comprising a mean of 139.2 electrode contacts per participant (range = 108–224). sEEG data were acquired using a standard clinical EEG system (EEG-1200C, Nihon Kohden Co.) at a sampling rate of 2,000 Hz. Stimulus-triggered electrical pulses were recorded alongside the sEEG data to enable precise synchronization with stimulus onset. All recordings were conducted at the patients' bedside.

Before electrode implantation, T1-weighted magnetic resonance imaging (MRI) using a 1.5-T scanner and computed tomography (CT) were obtained for each patient. Following implantation, an additional CT scan was acquired, skull-stripped, and co-registered to the preoperative anatomical MRI scan. Individual electrode contacts were then visually identified on the co-registered CT images and marked in each patient's native preoperative MRI space using Brainstorm. The preoperative structural MRI scans were subsequently processed in Brainstorm to segment and reconstruct each patient's cortical surface and to obtain the anatomical coordinates of each recording site in Montreal Neurological Institute (MNI) space. Figure 2-2a shows the electrode locations projected onto the brain surface, with colors indicating individual

participants. Across participants, recording sites were broadly distributed across cortical and subcortical regions, providing coverage of multiple large-scale brain systems relevant to memory and perceptual decision-making.

2.3.2. Preprocessing

sEEG data were preprocessed in MATLAB using the FieldTrip toolbox. Continuous signals were band-pass filtered between 0.5 and 250 Hz and downsampled from 2,000 to 500 Hz. Line noise and its harmonics were attenuated using band-stop filters centered at 50, 100, 150, and 200 Hz, each with a bandwidth of ±2 Hz. The signals were subsequently detrended and demeaned.

Given the susceptibility of clinical sEEG recordings to physiological, environmental, and recording-related artifacts, we applied a multistage artifact-rejection procedure combining automated metrics with detailed visual inspection. Artifacts were first identified through inspection of summary statistics, individual channels, and individual trials, and channels and trials containing prominent contamination were excluded. For participants whose recordings exhibited synchronous common-mode contamination, characterized by highly similar signals across many or all channels, the remaining signals were rereferenced to the common average across retained channels. When brief, isolated high-amplitude transients remained, segments exceeding a z-score threshold of 6 for at least 2 ms were identified, and the affected windows, including a 4-ms margin on either side, were replaced by linear interpolation. Finally, full-rank independent component analysis was performed using the extended Infomax algorithm. Components potentially reflecting non-neural artifacts were identified from multiple statistical and spectral characteristics—including kurtosis, high-frequency power, line-noise prominence,

variance, and spectral entropy—and were visually reviewed before removal. This combination of channel-, trial-, segment-, and component-level procedures was intended to minimize residual artifacts while preserving as much usable neural data as possible.

2.3.3. Time-Frequency Analysis

The cleaned sEEG data were epoched separately to the video-viewing period, the delay period, and the picture-presentation period extending until the participant's response. These epochs are hereafter referred to as the **video period**, **delay period**, and **decision period**, respectively.

Time–frequency power was estimated for each trial and electrode using the multitaper convolution method. Discrete prolate spheroidal sequence tapers were applied using a fixed 200-ms sliding window. The multitaper approach reduces spectral leakage and provides relatively stable power estimates, particularly at higher frequencies. The fixed window length also allowed temporal resolution to remain consistent across frequencies. Power was calculated from 1 to 250 Hz in 2-Hz increments, providing broad coverage of low-frequency and high-frequency activity. A spectral smoothing parameter of 10 Hz was used to improve the robustness of power estimates, although at the cost of reduced frequency specificity. Power was sampled at 50-ms intervals, allowing the temporal evolution of neural activity to be tracked with substantial overlap between successive windows. Data were zero-padded to the next power of two to facilitate efficient spectral computation, and trial-level power estimates were retained to support subsequent condition-specific and trial-level analyses.

Power values were normalized separately for each trial using z-score baseline correction. For all event-aligned analyses, the baseline was defined as the 1-s interval occurring 3 to 2 s before the

participant initiated the trial. This interval was selected to provide a relatively uncontaminated estimate of resting neural activity while avoiding motor preparatory signals that might arise immediately before the self-paced button press used to begin the trial. This trial-specific normalization reduced the influence of between-trial differences in overall power and slow fluctuations in signal amplitude, allowing event-related changes to be expressed relative to each trial's own pretrial activity.

2.4. Statistical Analysis

2.4.1. Univariate electrode-level analysis

All electrophysiological statistical analyses focused on *high-gamma* power between 70 and 150 Hz. Task-related broadband activity within this frequency range is widely used in intracranial electrophysiology as a sensitive and spatially localized index of cortical engagement. High-gamma power is closely associated with aggregate neuronal firing and can track rapid changes in local cortical processing across a range of perceptual and cognitive tasks [34–38]. Its combination of anatomical specificity and temporal precision therefore made it well suited for examining the evolving neural representations of memory–sensory correspondence and perceptual identity in the present study.

Univariate analyses were conducted separately for each participant. We first identified electrodes exhibiting significant event-related increases in high-gamma activity during the video, delay, and decision periods. For each period, trial-level high-gamma power was compared with power from a pretrial baseline beginning 3 s before the participant initiated the trial. Baseline and active

intervals were matched in duration; for the video period, a 2-s baseline segment ending 1 s before trial initiation was duplicated to match the 4-s active interval. This baseline placement was chosen to reduce contamination from motor preparation immediately preceding the self-paced start-button press. Active and baseline measurements from the same trials were compared using paired-sample cluster-based Monte Carlo permutation tests. To identify event-related increases, samples showing greater high-gamma power during the active period than during baseline at an initial threshold of $p < .025$ were grouped into clusters across contiguous time–frequency points. Positive clusters were retained at a cluster-level threshold of $p < .05$ based on 500 randomizations. Electrodes contributing to these clusters were classified as task-responsive for that period and retained for the corresponding condition-level analyses.

Within task-responsive electrodes identified for the decision period, we separately tested three contrasts designed to capture distinct forms of information hypothesized to contribute to perceptual identification. The **video–picture match contrast** compared pictures drawn from the preceding video with pictures drawn from a different video. Because the critical difference between these conditions was whether the newly presented sensory input corresponded to the representation maintained from the preceding video, this contrast was used to index *memory–sensory comparison*: neural activity sensitive to the correspondence between internally maintained contextual information and the incoming visual stimulus. The **picture-category contrast** compared pictures containing a primate with those containing a non-primate animal and was used to identify neural activity carrying information about the perceptual identity of the picture itself. Whereas the video–picture match contrast therefore indexed the relationship between internal and external representations, the picture-category contrast indexed the emerging representation of what was currently being perceived. Finally, the **decision-accuracy contrast**

compared trials on which the participant's perceptual judgment was correct with those on which it was incorrect. This contrast was used to capture neural information related to the strength or sufficiency of the evidence available for the perceptual decision—specifically, whether the available evidence was sufficient to support a correct judgment. Together, the three contrasts allowed us to examine where and when neural activity represented memory–sensory correspondence, perceptual identity, and decision-relevant evidence sufficient for successful perceptual judgments.

2.4.2. Multivariate decoding analysis

Multivariate pattern analyses were conducted separately for each participant across the video, delay, and decision periods to test whether distributed high-gamma activity encoded the same three trial dimensions examined in the univariate analyses: video–picture match, picture category, and decision accuracy. The video and delay periods were included as temporal control periods because the information defining these contrasts depended on the subsequently presented picture and perceptual response and therefore was not expected to be decodable before picture onset. Decoding during these periods thus provided a check against systematic pre-picture differences or other potential confounds. For all three periods, analyses were restricted to electrodes that showed significant task-related increases in high-gamma activity during the decision period relative to baseline, thereby using a common set of task-responsive electrodes across periods. High-gamma power between 70 and 150 Hz was analyzed over the corresponding task interval for each period: during video viewing, throughout the delay, and from picture onset through the participant-specific mean response interval during the decision period.

Decoding was performed using 300-ms sliding windows advanced in 50-ms increments. Within each window, high-gamma power was averaged across time and grouped into equally spaced frequency bins. Power values from all retained electrodes and frequency bins were then concatenated to form a multivariate feature vector for each trial. The number of frequency bins and cross-validation folds was adjusted conservatively according to the number of trials in the smaller class: analyses used two, three, or five folds and between four and ten frequency bins, depending on the available data. Contrasts with fewer than five trials in either class were not analyzed. This adaptive procedure limited model complexity when trial counts were low while allowing richer feature representations when more data were available.

At each time window, a linear support vector machine was trained and evaluated using repeated stratified *k*-fold cross-validation. Features were standardized using the mean and standard deviation estimated exclusively from the training fold. Principal component analysis was likewise fitted only to the training data, and the resulting transformation was subsequently applied to the held-out test data, thereby preventing information leakage. The number of retained components was chosen to explain up to 90% of the training-set variance but was conservatively capped according to the number of training observations and trials in the minority class. Inverse-frequency class weights were applied during model fitting, and performance was quantified using balanced accuracy to reduce sensitivity to unequal class sizes. Observed decoding accuracy was averaged across 100 repeated cross-validation partitions.

Statistical significance was evaluated using permutation testing with cluster-based correction across time. For each contrast, trial labels were randomly permuted 1,000 times, and decoding was repeated using 10 independently stratified cross-validation partitions for each permutation.

Observed accuracy at each window was standardized relative to its permutation distribution. Temporally contiguous windows exceeding a one-sided pointwise threshold of $p < .05$ were grouped into clusters, and cluster mass was defined as the sum of the standardized decoding values within each cluster. Cluster-level significance was determined by comparing the observed cluster mass with the permutation distribution of the maximum cluster mass, using a corrected threshold of $p < .05$. This procedure identified periods during which distributed high-gamma patterns reliably distinguished the trial classes while controlling for multiple comparisons across time.

Group-level inference was subsequently performed separately for each decoding contrast to determine whether classification accuracy was reliably greater than chance across participants. Because the duration of the decision period varied across participants, each participant's balanced-accuracy time course was linearly interpolated onto a common 50-ms time grid restricted to the temporal interval shared by all included participants. At each time point, balanced accuracy minus chance performance (0.50) was compared with zero using a one-tailed, dependent-samples *t* test. Statistical significance was assessed using cluster-based Monte Carlo permutation testing with an initial pointwise threshold of $p < .05$ and a cluster-level threshold of $p < .05$ based on 10,000 randomizations. Temporally contiguous samples exceeding the pointwise threshold were grouped into clusters, and cluster mass was defined as the sum of the *t* statistics within each cluster. Observed clusters were evaluated against the permutation distribution of the maximum cluster mass, thereby controlling for multiple comparisons across time. This analysis identified time periods during which decoding performance was consistently above chance across participants.

To identify the recording sites contributing to significant multivariate decoding, we performed a Haufe-pattern traceback analysis, which transforms discriminative classifier weights into activation patterns that are more interpretable in terms of the neural signals contributing to classification [39]. The analysis was performed separately for each group-level temporal cluster with a cluster-corrected value of $p \leq .05$. For each cluster, the search interval was extended by two 50-ms bins on either side and mapped onto each participant's decoding time course. Participants were included in the traceback when their mean decoding performance within this interval exceeded their subject-specific permutation distribution by a mean z score greater than 0.5. For each included participant, the time point showing the highest decoding accuracy within the search interval was identified, and the corresponding 300-ms window together with the immediately preceding and following windows was analyzed.

The original multivariate feature construction and adaptive cross-validation settings were retained. To reduce the influence of class imbalance, trials from the two classes were randomly subsampled to equal numbers in each of 50 repetitions. Within each training fold, a linear support vector machine was fitted following training-set standardization and principal component analysis. The classifier weights were transformed into activation patterns using the covariance structure of the training data and projected from principal-component space back into the original electrode-by-frequency feature space. The absolute magnitude of each activation-pattern value was then averaged across frequency bins, cross-validation folds, selected temporal windows, and repetitions to obtain a contribution estimate for each electrode. These absolute Haufe-pattern contributions were used to characterize the relative involvement and spatial distribution of electrodes supporting decoding, rather than as independent electrode-level significance tests.

2.4.3. Classifier-Evidence Analysis

Building on the significant group-level decoding clusters identified in the preceding multivariate analysis, we next examined whether trial-by-trial fluctuations in the strength of classifier evidence were related across clusters. For each significant cluster, we therefore extracted classifier decision values as a continuous measure of the strength of class-related neural information on individual trials, extending beyond the binary correct-or-incorrect output used to calculate decoding accuracy [40–42]. Related single-trial discriminant measures have been used to characterize variability in neural information associated with perceptual decisions [19, 43]. Group-defined temporal clusters were mapped onto the corresponding decoding windows for each participant, and the same feature-construction, dimensionality-reduction, and cross-validation procedures used in the decoding analyses were retained. Classifier decision values were obtained exclusively from held-out trials. Within each cross-validation fold, test-set decision values were standardized relative to the mean and standard deviation of the training-set decision values. The standardized held-out values were then averaged across the temporal windows belonging to each cluster and across 100 repeated cross-validation procedures. On each repetition, the two trial classes were randomly subsampled to equal sizes to minimize the influence of class imbalance. The absolute value of the resulting classifier evidence was used, such that larger values indicated stronger class-related neural information regardless of the class favored by the classifier.

Regression analyses were then used to test hypothesized trial-by-trial relationships among the absolute classifier-evidence measures obtained from different decoding clusters. Cluster-specific general high-gamma power was included as a covariate to reduce the possibility that

relationships between classifier-evidence measures reflected nonspecific fluctuations in overall neural activity. All continuous variables were standardized within participant before pooling trials across participants and fitting the models. For analyses involving interactions between classifier-evidence measures, the constituent variables were standardized before constructing the interaction term.

Population-level inference was performed using trial-level linear mixed-effects models fitted by restricted maximum likelihood. Each model included the classifier-evidence variables required to test the hypothesized relationship together with the corresponding high-gamma power covariates. For interaction analyses, both constituent main effects and their interaction were included in the model. To account for repeated observations within participants while allowing the hypothesized relationship to vary across individuals, each model included a participant-specific random intercept and a participant-specific random slope for the fixed effect of interest. Random intercepts and slopes were specified as separate random-effects terms and were therefore estimated without an intercept–slope covariance. The prespecified fixed effects of interest were evaluated using two-tailed tests.

3. Results

3.1. Behavioral Results

Before analysis, trials with response times below 300 ms or above 4,000 ms, or with an absolute within-participant robust MAD score greater than 3.5, were excluded from all behavioral

analyses. Participants performed the perceptual decision task with relatively high accuracy overall (M = .85, SD = .08), with a mean response time of 2.18 s (SD = 0.51 s). Accuracy did not differ significantly between primate and non-primate pictures (.83 vs. .86).

Perceptual sensitivity (d') was significantly greater when the picture was drawn from the preceding video than when it was drawn from a different video (paired $t(9) = 4.60$, $p < .01$; Figure 2-1b). Response bias was similar between the same-video and different-video conditions (.13 vs. .09), indicating that the sensitivity advantage was not attributable to a shift in response tendency.

Because this effect could also reflect a match between the categories of the video and picture, we repeated the analysis while holding video–picture category correspondence constant. Specifically, pictures drawn from the preceding video were compared with pictures drawn from a different video containing the same animal category. Perceptual sensitivity remained significantly greater in the same-video condition (3.10 vs. 2.38; paired $t(9) = 3.03$, $p < .05$). Similarly, for primate pictures, accuracy was significantly greater when the picture was drawn from the same primate video than from a different primate video (.91 vs. .80; paired $t(9) = 2.97$, $p < .05$). Together, these results indicate that perceptual sensitivity was enhanced by correspondence with the specific content of the preceding video, beyond a general match in animal category.

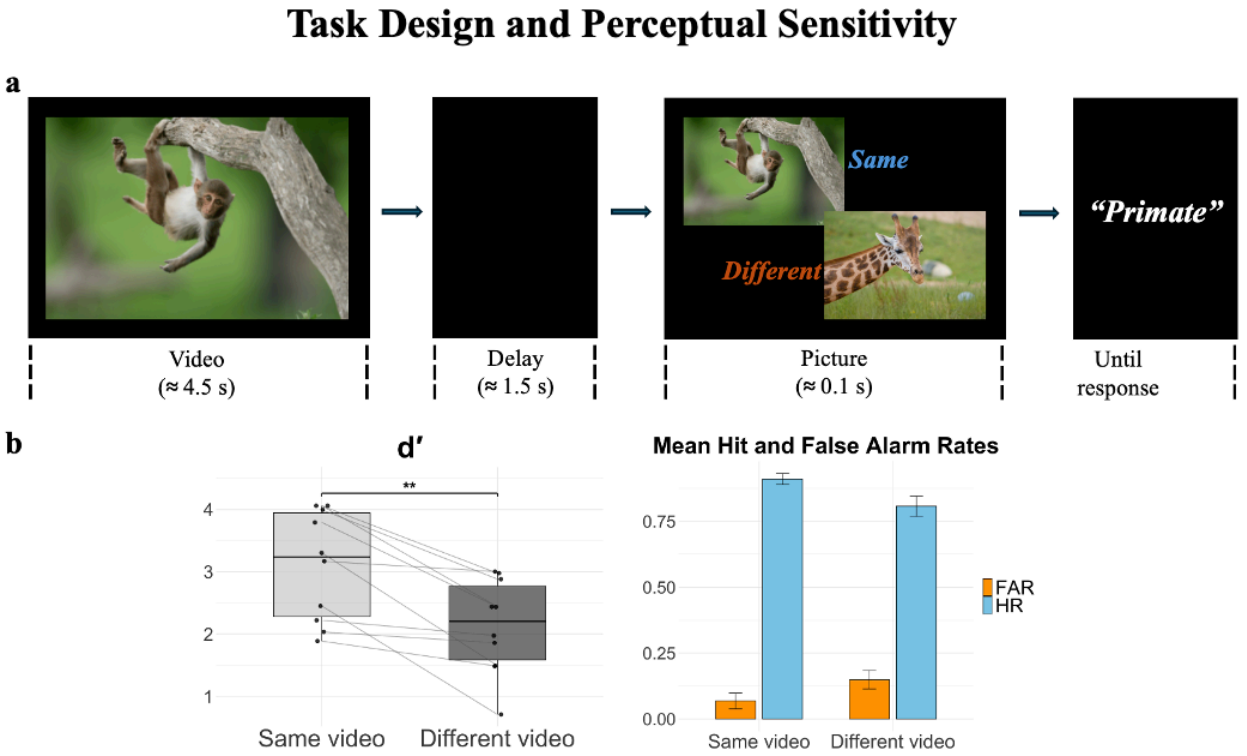

Figure 2-1. (a) Schematic of the perceptual detection task. Participants viewed a trial-unique 4–5-s video depicting animals, followed by an approximately 1.5-s blank delay. A picture was then presented for approximately 100 ms, after which participants indicated as quickly as possible whether the picture contained a primate. **(b)** Behavioral performance as a function of whether the picture was drawn from the preceding video (same video) or from a different video. Left, perceptual sensitivity (d') was significantly greater in the same-video condition than in the different-video condition (paired $t(9) = 4.60$, $p < .01$). Right, hit rate and false-alarm rate for the two conditions; the same-video condition showed a higher hit rate and a lower false-alarm rate than the different-video condition.

3.2. Univariate Results

Having established that perceptual sensitivity was enhanced when the picture matched the preceding video, we next examined intracranial activity across different periods of the task by identifying electrodes showing significant event-related increases in high-gamma activity relative to the pretrial baseline (cluster-corrected $p < .05$; Figure 2-2b). Task-responsive electrodes were observed during all three task periods, with their spatial distribution becoming

progressively broader across the trial. During the video period, responsive electrodes were concentrated predominantly in posterior occipital and occipitotemporal regions, consistent with sustained visual processing of the video. During the delay period, responsive electrodes remained prominent in posterior regions but extended more broadly into temporal and parietal association cortex. During the decision period, responsive electrodes were distributed more widely across the brain, spanning posterior sensory and association regions and extending into frontal cortex, consistent with the additional cognitive and decision-related processes engaged following picture presentation.

Within the task-responsive electrodes identified for the decision period, we next tested three contrasts intended to capture successive forms of information relevant to perceptual identification: video–picture match, indexing correspondence between the internally maintained video representation and the incoming sensory input; picture category, indexing the perceptual identity of the presented picture; and decision accuracy, indexing decision-relevant evidence sufficient to support a correct perceptual judgment. Electrodes showing significant condition effects are indicated in Figure 2-2b by white outlines, with black arrows marking electrodes significant for the video–picture match contrast and gray arrows marking those significant for the picture-category contrast. The video–picture match contrast yielded five significant electrodes across four participants (cluster-corrected $p < .05$), several of which converged across participants in posterior medial and parietal regions encompassing DMN-associated precuneus territory, with an additional frontal electrode. This convergence is notable given the proposed role of the precuneus as a posterior DMN hub positioned to coordinate internally and externally oriented processing. Recent work has shown that precuneus subdivisions flexibly interact with networks associated with both internal and external processing and has specifically proposed that

this architecture enables stored internal representations to contribute to the moment-to-moment interpretation of external information [32, 44]. Thus, although spatially limited, the video–picture match effect was consistent with our hypothesis that DMN-associated regions provide an internal contextual representation against which newly arriving sensory information can be evaluated. The picture-category contrast yielded four significant electrodes across three participants (cluster-corrected $p < .05$), located primarily in posterior and ventral temporal regions. This distribution was likewise consistent with the contrast's intended role in capturing perceptual identity, as posterior and ventral temporal cortex contains representations that differentiate visual object categories [45]. No electrodes survived correction for the decision-accuracy contrast.

These univariate results therefore provided preliminary anatomical support for the distinction among memory–sensory correspondence, perceptual identity, and decision-relevant evidence, but the effects were sparse and captured only a small fraction of the task-responsive electrodes. This limited sensitivity is not unexpected: condition information need not appear as a large difference in mean activity at any single recording site, but can instead be carried by the joint configuration of weaker signals distributed across multiple sites. In the present sEEG data, this limitation is compounded by sparse and participant-specific anatomical sampling and by the stringent correction required for electrode-wise time–frequency comparisons. Consequently, the absence of a strong univariate effect cannot be taken to indicate that the corresponding information was not represented in the recorded population. We therefore next used multivariate pattern analysis to test whether the three forms of information could be recovered from distributed patterns of high-gamma activity, allowing information expressed collectively across

electrodes to be detected even when individual electrodes showed weak or inconsistent condition effects.

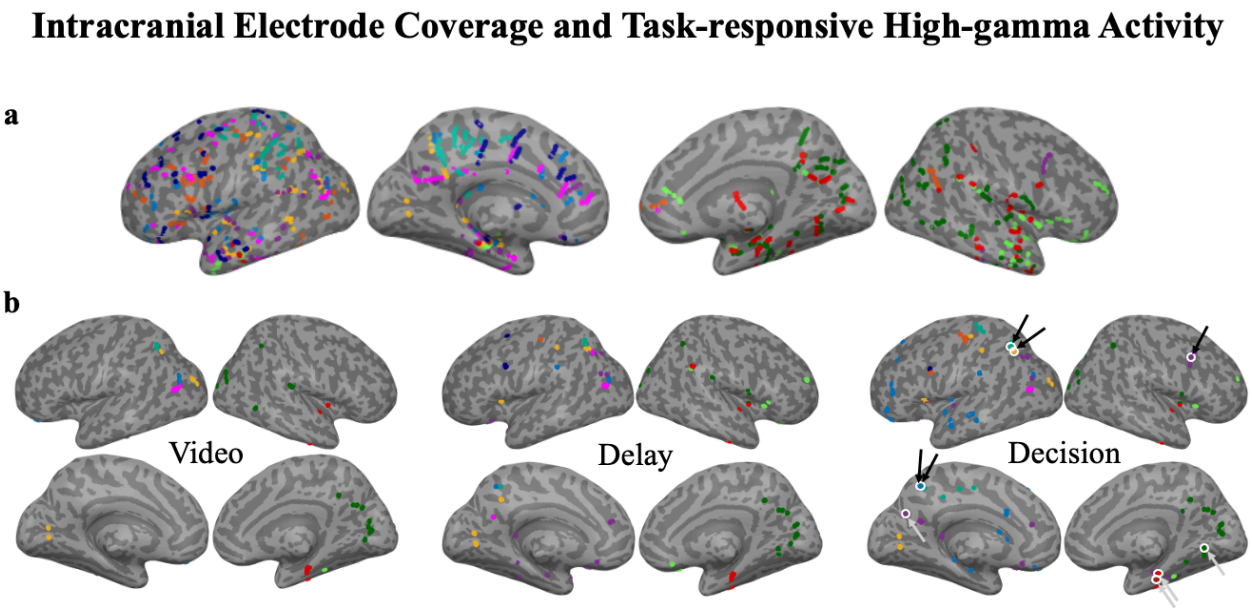

Figure 2-2. (a) Spatial distribution of all intracranial electrodes included in the study, projected onto a common brain surface. **(b)** Electrodes showing significant event-related increases in high-gamma activity relative to the pretrial baseline during the video, delay, and decision periods (cluster-corrected $p < .05$). During the video period, task-responsive electrodes were concentrated predominantly in posterior visual regions; their distribution extended more broadly during the delay period and became widespread during the decision period, including frontal regions. Within decision-period task-responsive electrodes, white outlines indicate electrodes showing significant differences between experimental conditions (cluster-corrected $p < .05$). Black arrows indicate electrodes significant for the video–picture match contrast, and gray arrows indicate electrodes significant for the picture-category contrast. No electrodes survived correction for the decision-accuracy contrast.

3.3. Multivariate Decoding Results

We next performed multivariate decoding analyses separately for each participant and subsequently conducted group-level inference to determine whether the three trial dimensions—video–picture match, picture category, and decision accuracy—could be reliably decoded across participants. Decoding was evaluated using repeated cross-validated linear support vector

machines with permutation-based significance testing and cluster-based correction across time. All three contrasts showed significant above-chance decoding during the decision period, indicating that distributed high-gamma activity reliably distinguished the corresponding trial types (Figure 2-3, left). In contrast, no significant decoding clusters were observed during the video or delay control periods, confirming that successful decoding during the decision period was not attributable to systematic differences already present before picture onset. To characterize the spatial distribution of the signals supporting these effects, we performed a Haufe-pattern traceback analysis for each significant decoding cluster. Figure 2-3 (right) shows the electrodes contributing most strongly to each cluster among participants with reliable subject-level decoding, with electrodes associated with the first and second significant clusters of the video–picture match and picture-category contrasts distinguished by white and black outlines, respectively.

For the video–picture match contrast, decoding was already elevated immediately after picture presentation and remained above chance across much of the early-to-middle decision period before gradually returning toward chance level (Figure 2-3, top left). Although the cluster-based analysis formally identified two significant clusters (Cluster 1: 0.42–0.72 s; Cluster 2: 0.87–1.12 s; both cluster-corrected $p < .05$), they were separated by only a brief nonsignificant interval and merged at a more permissive statistical threshold (cluster-corrected $p < .10$). Together, this pattern suggests that information about video–picture correspondence arose early and was represented in a largely sustained manner following picture onset, consistent with our hypothesis that comparison between working-memory and sensory representations would emerge rapidly and persist over time. The Haufe-pattern traceback analysis indicated that electrodes contributing to this decoding were distributed predominantly across DMN-associated regions, with additional

contributions from a smaller number of frontoparietal and visual sites (Figure 2-3, top right). This spatial pattern is consistent with the proposed role of the DMN in maintaining representations beyond the immediate sensory moment and using internally retained information to guide ongoing processing. In the present task, such temporally extended representations could allow information from the preceding video to remain available while the brief picture continued to be processed, supporting a sustained comparison between working memory and the incoming sensory representation even after the picture had disappeared.

The picture-category contrast showed a different temporal profile. Decoding began near chance level after picture onset and then increased sharply, producing an early significant cluster (Cluster 1: 0.52–0.72 s; cluster-corrected $p < .05$; Figure 2-3, middle left). Following an intervening decline, decoding increased again to form a distinct later cluster (Cluster 2: 1.07–1.42 s; cluster-corrected $p < .05$) and remained strongly elevated toward the end of the analyzed decision period. Thus, unlike the relatively sustained video–picture match signal, picture-category information appeared to develop after picture onset and showed two temporally separated periods of reliable representation. The Haufe-pattern traceback analysis showed that electrodes contributing to picture-category decoding were distributed more prominently across visual and sensorimotor-associated regions, with relatively fewer sites in DMN- or frontoparietal-associated regions than for the video–picture match contrast (Figure 2-3, middle right). This distribution is consistent with the picture-category signal reflecting developing visual recognition of the presented picture, while involvement of sensorimotor regions may additionally reflect the transformation of perceptual information into the button-press response required by the task.

The decision-accuracy contrast likewise began near chance level and increased later in the decision period, consistent with the gradual emergence of neural information related to whether sufficient evidence had accumulated to support a correct perceptual judgment (Figure 2-3, bottom left). Its significant cluster (0.77–0.97 s; cluster-corrected $p < .05$) emerged after the initial rise in picture-category decoding and fell temporally between the two significant picture-category clusters, while overlapping with the later significant period of video–picture match decoding. The Haufe-pattern traceback analysis revealed a broadly distributed set of contributing electrodes spanning frontoparietal, DMN-associated, visual, and sensorimotor regions, with relatively greater frontoparietal involvement than observed for the video–picture match or picture-category contrasts (Figure 2-3, bottom right). Such broad recruitment is consistent with decision-evidence formation requiring the integration of information across multiple systems, while the stronger frontoparietal contribution accords with the proposed role of these regions in representing and integrating task-relevant evidence toward an impending choice. The temporal position of this signal, between the two periods of picture-category decoding and overlapping with the later comparison signal, further suggests that it may reflect an intermediate stage in which information generated by memory–sensory comparison is integrated into evidence supporting perceptual identification.

Taken together, the temporal profiles across the three contrasts suggest that the represented information may not develop independently, but instead evolve in a coordinated manner across the decision period. The early emergence of video–picture match information suggests that memory–sensory comparison begins rapidly and contributes to the initial development of picture-category information. As this comparison remains active, it may continue to interact with the emerging picture representation, allowing additional decision-relevant evidence to develop

over time. The later re-emergence of strong picture-category information is therefore consistent with a process in which an initial perceptual judgment is progressively refined by decision-relevant evidence generated through continued comparison between the sensory representation and information maintained in working memory. We next tested these proposed relationships directly by examining trial-by-trial covariation in classifier evidence across the significant decoding clusters.

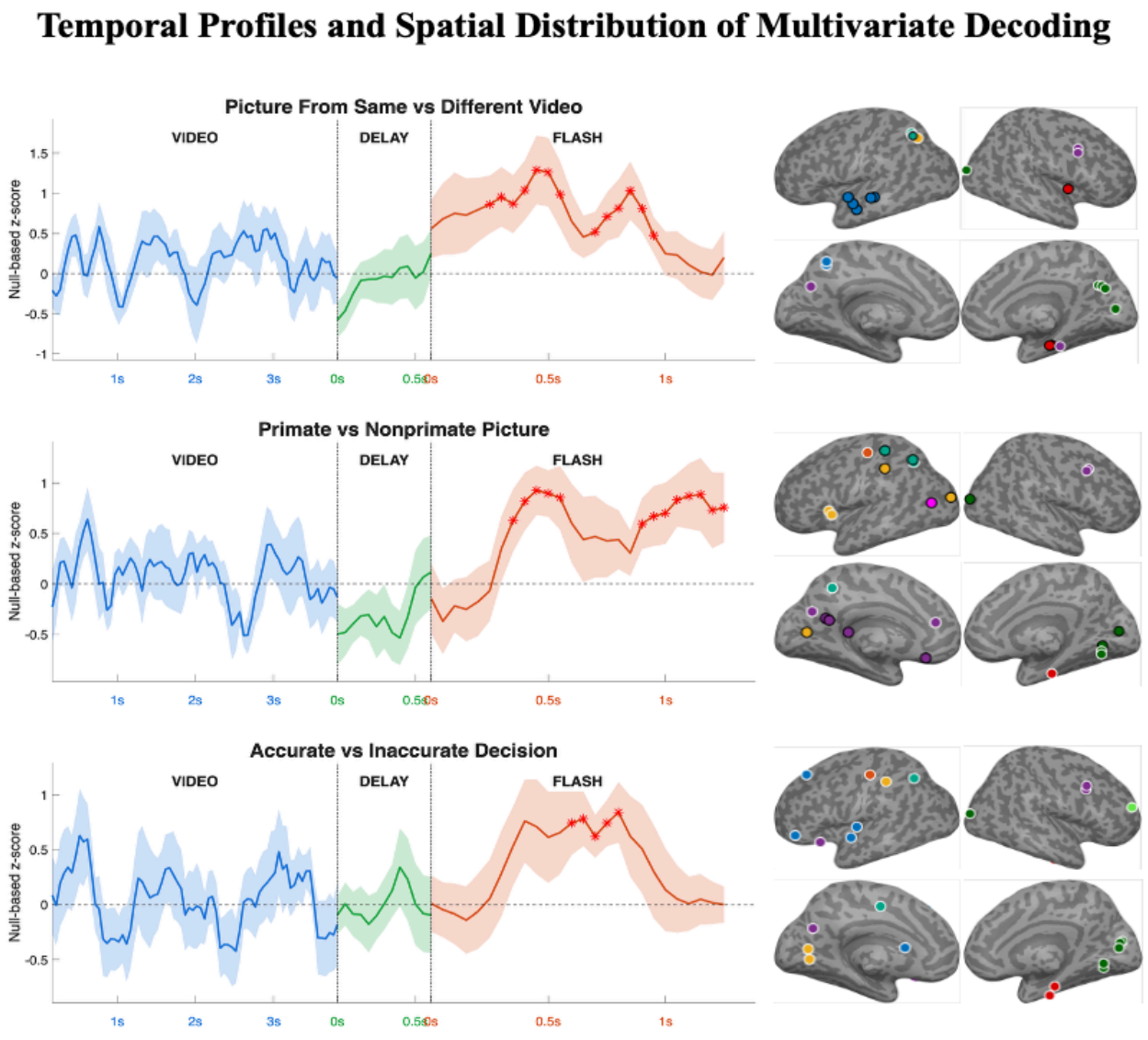

Figure 2-3. Left: Group-level decoding time courses for video–picture match (top), picture category (middle), and decision accuracy (bottom). Decoding performance for each participant was standardized relative to the corresponding permutation-based null distribution, such that zero indicates chance-level decoding. Lines show the group mean and shaded regions indicate ± SEM. Asterisks indicate time periods belonging to significant group-level clusters ($p < .05$, cluster-corrected). The video and delay periods served as temporal control periods and are displayed with compressed horizontal scaling for illustrative purposes; the underlying time points were not removed or averaged. Because decision-period duration varied across participants, the group-level time course was restricted to the range of decoding windows shared by all participants. Accordingly, the end of the displayed decision-period axis reflects the last common decoding window, determined by the participant with the shortest analyzed response interval, rather than the overall mean response time (mean RT = 2.18 s). **Right:** Spatial distribution of electrodes contributing to significant multivariate decoding clusters, identified using the Haufe-pattern traceback analysis. Displayed electrodes were restricted to participants with a mean subject-level decoding z score > 0.5 within the corresponding cluster and to electrodes with an absolute within-subject Haufe-pattern z score > 1. For the video–picture match and picture-category contrasts, which each contained two significant temporal clusters, electrodes associated with the first cluster are outlined in white and those associated with the second cluster are outlined in black.

3.4. Classifier-Evidence Analysis Results

The temporal ordering of the decoding effects suggested that memory–sensory comparison, picture-category information, and decision-related evidence may be dynamically related across the decision period. To examine these relationships at the trial level, we tested whether fluctuations in absolute classifier evidence were related across the significant decoding clusters while controlling for cluster-specific general high-gamma power (Figure 2-4).

We first asked whether early memory–sensory comparison was related to the initial development of picture-category information. Classifier evidence in the first video–picture match cluster positively predicted classifier evidence in the first picture-category cluster ($\beta = 0.092$, $t(1457) = 2.01$, $p < .05$; Figure 2-4, left). Thus, on trials in which the neural distinction between matching and nonmatching pictures was stronger early in the decision period, the subsequent neural representation distinguishing picture category was also stronger.

We next tested whether the later video–picture comparison signal interacted with the emerging picture-category representation in predicting decision-accuracy evidence. A significant positive

interaction was observed between classifier evidence in the first picture-category cluster and the second video–picture match cluster in predicting classifier evidence in the decision-accuracy cluster ($\beta = 0.129$, $t(1454) = 2.61$, $p < .01$; Figure 2-4, middle). At the mean level of video–picture match Cluster 2 evidence, picture-category Cluster 1 evidence also positively predicted decision-accuracy evidence ($\beta = 0.052$, $t(1454) = 2.06$, $p < .05$), whereas the conditional main effect of video–picture match Cluster 2 evidence, evaluated at the mean level of picture-category Cluster 1 evidence, was not significant ($\beta = 0.006$, $t(1454) = 0.25$). Critically, the interaction indicates that the relationship between early picture-category information and subsequent decision-accuracy evidence became increasingly positive as later memory–sensory comparison evidence increased. Given that the decision-accuracy contrast indexes whether the available information was sufficient to support a correct perceptual judgment, this pattern is consistent with the contribution of emerging picture-category information to decision-related evidence depending on the strength of continued comparison with the representation maintained from the preceding video.

Finally, classifier evidence in the decision-accuracy cluster positively predicted classifier evidence in the second picture-category cluster ($\beta = 0.067$, $t(1457) = 2.49$, $p < .05$; Figure 2-4, right). Trials containing stronger decision-related evidence were therefore followed by stronger later picture-category information.

Together, the three relationships suggest a temporally ordered process in which early memory–sensory comparison is associated with the initial development of perceptual identification; the emerging picture-category representation and continuing memory–sensory comparison jointly predict the strength of subsequent decision-related evidence; and this decision-related evidence, in turn, predicts the later representation of picture category. This pattern is consistent with an

iterative process in which comparison between working memory and sensory information continues to contribute to the development of evidence supporting perceptual identification over time.

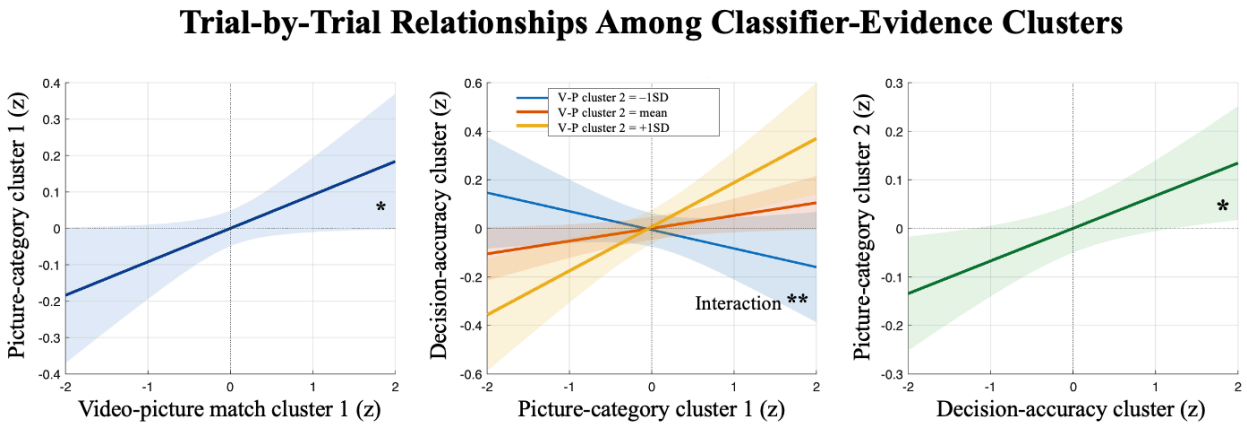

Figure 2-4. Population-level predictions from linear mixed-effects models relating absolute classifier evidence across significant decoding clusters. All classifier-evidence measures were standardized within participant, and cluster-specific general high-gamma power covariates were held at their within-participant means. **Left:** Classifier evidence in video–picture match Cluster 1 positively predicted classifier evidence in picture-category Cluster 1 ($\beta = 0.092$, $p < .05$). **Middle:** Picture-category Cluster 1 evidence interacted positively with video–picture match Cluster 2 evidence in predicting decision-accuracy Cluster 1 evidence ($\beta = 0.129$, $p < .01$). Predicted relationships are shown at video–picture match Cluster 2 evidence values of −1 SD, the within-participant mean, and +1 SD. **Right:** Decision-accuracy Cluster 1 evidence positively predicted classifier evidence in picture-category Cluster 2 ($\beta = 0.067$, $p < .05$). Lines show population-level fixed-effect predictions, and shaded regions indicate 95% confidence intervals.

4. Discussion

Perceiving a briefly glimpsed stimulus in light of the recent past requires the brain to bring into register two representations that differ in origin, format, and timing: an internally maintained trace of prior experience and a transient sensory input. The present findings characterize how this registration is achieved, suggesting that memory–sensory comparison operates not as a discrete

matching operation but as a sustained computation that is anatomically dissociable from, yet dynamically coupled to, the formation of perceptual and decision-related representations. In what follows, we consider what this temporally extended architecture implies for accounts of attentional templates and predictive processing, for the functional characterization of the default mode and frontoparietal networks, and for models of how perceptual decisions are constructed from multiple sources of evidence.

A first implication concerns how memory-based facilitation of perception should be conceptualized. Influential template-based accounts, rooted in the biased-competition framework, propose that actively maintained content pre-activates matching sensory representations so that corresponding input gains a competitive advantage [46, 47]. These accounts explain why matching stimuli are prioritized, but they are largely agnostic about the temporal structure of the matching process itself, which is often implicitly treated as instantaneous or as a preparatory state established before stimulus onset. The present observation that correspondence information emerged rapidly and then persisted across much of the decision period—continuing to modulate the translation of perceptual identity into decision-relevant evidence—suggests instead that the comparison behaves like an ongoing source of evidence in its own right. Within sequential-sampling frameworks, a decision variable evolves by integrating any signal that systematically discriminates the response alternatives [48, 49]. Our results suggest that the output of a memory–sensory comparison can constitute such a signal: rather than merely shifting a starting point or lowering a threshold prior to sampling, correspondence with working memory appears to be consulted repeatedly while evidence is being formed. This offers a mechanistic complement to behavioral modeling work on memory-driven detection advantages

and suggests that formal models of memory-guided perception may need to incorporate internally generated, temporally extended evidence sources alongside stimulus-driven ones.

The findings also bear on predictive-processing accounts, which hold that perception arises from the reconciliation of top-down predictions with bottom-up input [50–52]. Empirically, expectations have been shown to sharpen sensory representations [53] and to instantiate stimulus templates in advance of sensory input [54], and memory systems have been proposed to supply such predictions, for example through hippocampal pattern completion feeding expected content into visual cortex [55, 56]. In much of this literature, however, the moment of comparison is conceived as a rapid registration of match or mismatch at stimulus onset, akin to the transient comparator responses observed in the hippocampus when unfolding events violate stored associations [57, 58]. The present data suggest that when the internal representation is a rich, trial-unique episode and the external input is too brief to be fully resolved on arrival, comparison takes a different form: correspondence information was sustained for roughly a second after stimulus offset and interacted with concurrently developing perceptual representations. This pattern is more consistent with an iterative interrogation of the sensory trace against memory than with a single prediction-error computation, and it suggests that the timescale of memory–sensory comparison may adapt to the resolvability of the input—a possibility that static match–mismatch paradigms are not positioned to reveal.

The anatomical distribution of the comparison signal speaks to an evolving understanding of the default mode network. Long characterized as task-negative and oriented away from the external environment [59, 60], the DMN has more recently been situated at the apex of a cortical hierarchy—maximally distant from primary sensory regions along the principal gradient of

cortical organization [61, 62]—with correspondingly long intrinsic timescales suited to integrating information across extended windows [63–65]. Nearly all evidence for this integrative role, however, derives from fMRI during naturalistic or narrative paradigms, leaving open whether the DMN contributes to cognition on the timescale of individual perceptual decisions. The present intracranial results indicate that DMN-associated regions carry trial-specific information about the relation between remembered and current input within hundreds of milliseconds of stimulus onset, and that fluctuations in this information predict the downstream development of perceptual and decision signals. This pattern is consistent with a view of the DMN not as a system that disengages when the external world demands attention, but as a functional interface at which retained context is actively brought to bear on incoming input—in line with hierarchical accounts of its connectivity, yet operating at a speed those accounts have rarely been able to test.

The decision-accuracy contrast showed a complementary spatial pattern. Electrodes contributing to this signal were broadly distributed across frontoparietal, DMN-associated, visual, and sensorimotor regions, with frontoparietal sites contributing relatively more prominently than they did to the other contrasts. This relative emphasis accords with a broad literature implicating frontoparietal cortex in accumulating and representing information for an impending choice, from single-unit recordings in parietal cortex [66] to supramodal accumulation signals in human electrophysiology [67] and comparative syntheses across species [68], and with evidence that choice-related information emerges progressively in frontal and parietal regions as decisions form [69]. On this reading, the decision-relevant evidence observed here would have been informed not only by feedforward sensory input but also by the output of a comparison expressed more strongly in DMN-associated territory. Together with demonstrations that

frontoparietal mechanisms select and transform the contents of working memory itself [70], this pattern is suggestive of a graded division of labor in which contextual evaluation and evidence integration draw differentially on distinct large-scale systems. We emphasize, however, that the broad distribution of contributing sites argues against any strict anatomical segregation: what the data support is a difference in relative contribution across systems, not an exclusive localization of decision evidence to frontoparietal cortex, and coordination among systems, rather than the operation of any one system alone, is likely what supports the resulting judgment.

The trial-level dependency of later picture-category information on preceding decision-related evidence further suggests that perceptual representations continued to be shaped after an initial categorization had begun. This is difficult to reconcile with strictly feedforward models of recognition and instead accords with recurrent-processing accounts, in which perception of degraded or briefly presented input is completed through iterative feedback loops [71]. Causal and computational evidence indicates that recurrence is specifically required when recognition is challenging [72], and top-down facilitation models propose that coarse initial information rapidly generates candidate interpretations that are then tested against accumulating sensory detail [73]. The late re-emergence of category information observed here, following and predicted by decision-related evidence, may reflect precisely such a consolidation phase, in which a fragile trace of the 100-ms picture is progressively stabilized into a robust categorical representation adequate to guide the response. Viewed this way, the full sequence of relationships we observed echoes, at the level of distributed human cortical activity, the classical proposal that percepts are hypotheses refined against evidence rather than read out in a single pass.

Several limitations warrant consideration. First, the delay period was not directly probed in the present study. Although a subset of electrodes showed significant increases in high-gamma activity during the delay, distributed predominantly across occipital and posterior parietal regions as well as ventral and medial temporal sites (Figure 2-2), the present analyses were not designed to characterize the information represented during this period. This focus was partly intentional, as our primary aim was to characterize the information directly involved in the subsequent perceptual decision. Accordingly, the present data do not establish how, or in what format, the video representation was maintained across the delay—whether through persistent activity in these delay-active regions or through complementary mechanisms such as activity-silent or distributed coding [74, 75]. Directly characterizing this maintenance process remains an important question for future work. Second, the constraints of clinical intracranial research apply: electrode coverage was participant-specific and dictated by clinical considerations, and the sample was modest in size [76]. Network-level attributions based on this coverage are therefore provisional, and medial temporal structures central to comparator accounts of novelty and match detection were not systematically sampled, leaving the contribution of the hippocampus to the present comparison process an open question. Third, the classifier-evidence relationships are correlational; establishing that comparison-related activity causally drives the development of decision evidence will require perturbation approaches such as direct electrical stimulation of the implicated sites. Finally, and most importantly, our analyses were not a direct test of a memory-comparison, evidence-accumulation, and recurrent-refinement sequence per se. The temporal ordering of the decoding effects and the trial-level relationships among them are consistent with this conceptual reading of the observed dynamics, but they do not isolate it experimentally, and other processing schemes could in principle produce similar patterns. The

framework outlined in this Discussion should therefore be read as an interpretation suggested by the observed temporal dynamics—one whose value lies in organizing the findings and generating testable predictions—rather than as a demonstrated mechanism, and the implications drawn here should be received with corresponding caution.

Notwithstanding these limitations, the present findings suggest a candidate architecture for memory-guided perception: a comparison between internally maintained and incoming representations that begins early, persists over time, and is expressed most strongly in default-mode territory; a translation of its output into decision-relevant evidence expressed across widely distributed regions with relatively greater frontoparietal involvement; and an iterative refinement of the perceptual representation that continues until the response. More broadly, they suggest that even the perception of a stimulus glimpsed for a tenth of a second is not a feedforward reading of the retinal input but an extended transaction between what the brain retains and what the world provides—one in which the boundary between remembering and perceiving is a matter of ongoing negotiation rather than fixed division.

References

1. Gregory, R. L. (1980). Perceptions as hypotheses. Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 290, 181–197.
2. Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. W. H. Freeman.
3. Bar, M. (2004). Visual objects in context. Nature Reviews Neuroscience, 5, 617–629.
4. Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. Nature Reviews Neuroscience, 15, 745–756.
5. Fritsche, M., Mostert, P., & de Lange, F. P. (2017). Opposite effects of recent history on perception and decision. Current Biology, 27, 590–595.
6. Kiyonaga, A., & Egner, T. (2013). Working memory as internal attention: Toward an integrative account of internal and external selection processes. Psychonomic Bulletin & Review, 20, 228–242.
7. D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. Annual Review of Psychology, 66, 115–142.
8. Kang, M.-S., Hong, S. W., Blake, R., & Woodman, G. F. (2011). Visual working memory contaminates perception. Psychonomic Bulletin & Review, 18, 860–869.
9. Gayet, S., Paffen, C. L. E., & Van der Stigchel, S. (2013). Information matching the content of visual working memory is prioritized for conscious access. Psychological Science, 24, 2472–2480.
10. Gayet, S., van Maanen, L., Heilbron, M., Paffen, C. L. E., & Van der Stigchel, S. (2016). Visual input that matches the content of visual working memory requires less (not faster) evidence sampling to reach conscious access. Journal of Vision, 16, 26.

11. Teng, C., & Kravitz, D. J. (2019). Visual working memory directly alters perception. Nature Human Behaviour, 3, 827–836.

12. Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. Nature Reviews Neuroscience, 6, 97–107.

13. Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. Nature, 458, 632–635.

14. Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. Psychological Science, 20, 207–214.

15. Gayet, S., Guggenmos, M., Christophel, T. B., Haynes, J.-D., Paffen, C. L. E., Van der Stigchel, S., & Sterzer, P. (2017). Visual working memory enhances the neural response to matching visual input. Journal of Neuroscience, 37, 6638–6647.

16. Hyun, J., Woodman, G. F., Vogel, E. K., Hollingworth, A., & Luck, S. J. (2009). The comparison of visual working memory representations with perceptual inputs. Journal of Experimental Psychology: Human Perception and Performance, 35, 1140–1160.

17. Soto, D., & Humphreys, G. W. (2006). Seeing the content of the mind: Enhanced awareness through working memory in patients with visual extinction. Proceedings of the National Academy of Sciences of the United States of America, 103, 4789–4792.

18. Soto, D., Wriglesworth, A., Bahrami-Balani, A., & Humphreys, G. W. (2010). Working memory enhances visual perception: Evidence from signal detection analysis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 441–456.

19. Philiastides, M. G., & Sajda, P. (2006). Temporal characterization of the neural correlates of perceptual decision making in the human brain. Cerebral Cortex, 16, 509–518.

20. O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. Nature Neuroscience, 15, 1729–1735.

21. Gherman, S., Markowitz, N., Tostaeva, G., et al. (2024). Intracranial electroencephalography reveals effector-independent evidence accumulation dynamics in multiple human brain regions. Nature Human Behaviour, 8, 758–770.

22. Ploran, E. J., Nelson, S. M., Velanova, K., Donaldson, D. I., Petersen, S. E., & Wheeler, M. E. (2007). Evidence accumulation and the moment of recognition: Dissociating perceptual recognition processes using fMRI. Journal of Neuroscience, 27, 11912–11924.

23. Wilming, N., Murphy, P. R., Meyniel, F., & Donner, T. H. (2020). Large-scale dynamics of perceptual decision information across human cortex. Nature Communications, 11, 5109.

24. Pereira, M., Mégevand, P., Tan, M. X., et al. (2021). Evidence accumulation relates to perceptual consciousness and monitoring. Nature Communications, 12, 3261.

25. Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. F. (2008). The spatiotemporal profile of cortical processing leading up to visual perception. Journal of Vision, 8(1), 12.

26. Tang, H., Schrimpf, M., Lotter, W., et al. (2018). Recurrent computations for visual pattern completion. Proceedings of the National Academy of Sciences of the United States of America, 115, 8835–8840.

27. Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. Journal of Neuroscience, 31, 2906–2915.

28. Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. Nature Communications, 7, 12141.

29. Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. Nature Neuroscience, 20, 115–125.

30. Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. Neuron, 95, 709–721.e5.

31. Konishi, M., McLaren, D. G., Engen, H., & Smallwood, J. (2015). Shaped by the past: The default mode network supports cognition that is independent of immediate perceptual input. PLoS ONE, 10, e0132209.

32. Lyu, D., Stieger, J. R., Davey, C. G., et al. (2021). A precuneal causal loop mediates external and internal information integration in the human brain. Journal of Neuroscience, 41, 9944–9956.

33. Heekeren, H. R., Marrett, S., Bandettini, P. A., & Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. Nature, 431, 859–862.

34. Crone, N. E., Miglioretti, D. L., Gordon, B., & Lesser, R. P. (1998). Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. Brain, 121, 2301–2315.

35. Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J., & Hsiao, S. S. (2008). Neural correlates of high-gamma oscillations (60–200 Hz) in macaque local field potentials and their potential implications in electrocorticography. Journal of Neuroscience, 28, 11526–11536.

36. Miller, K. J., Schalk, G., Fetz, E. E., den Nijs, M., Ojemann, J. G., & Rao, R. P. N. (2010). Cortical activity during motor execution, motor imagery, and imagery-based online feedback. Proceedings of the National Academy of Sciences of the United States of America, 107, 4430–4435.

37. Ray, S., & Maunsell, J. H. R. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLoS Biology, 9, e1000610.

38. Miller, K. J., Honey, C. J., Hermes, D., Rao, R. P. N., den Nijs, M., & Ojemann, J. G. (2014). Broadband changes in the cortical surface potential track activation of functionally diverse neuronal populations. NeuroImage, 85, 711–720.

39. Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage, 87, 96–110.

40. Sajda, P., Philiastides, M. G., & Parra, L. C. (2009). Single-trial analysis of neuroimaging data: Inferring neural networks underlying perceptual decision-making in the human brain. IEEE Reviews in Biomedical Engineering, 2, 97–109.

41. Muraskin, J., Brown, T. R., Walz, J. M., et al. (2018). A multimodal encoding model applied to imaging decision-related neural cascades in the human brain. NeuroImage, 180, 211–222.

42. Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. NeuroImage, 173, 434–447.

43. Ratcliff, R., Philiastides, M. G., & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. Proceedings of the National Academy of Sciences of the United States of America, 106, 6539–6544.

44. Utevsky, A. V., Smith, D. V., & Huettel, S. A. (2014). Precuneus is a functional core of the default-mode network. Journal of Neuroscience, 34, 932–940.

45. Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nature Neuroscience, 2, 913–919.

46. Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. Nature, 363, 345–347.

47. Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. Annual Review of Neuroscience, 18, 193–222.

48. Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. Annual Review of Neuroscience, 30, 535–574.

49. Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. Neural Computation, 20, 873–922.

50. Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2, 79–87.

51. Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11, 127–138.

52. de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? Trends in Cognitive Sciences, 22, 764–779.

53. Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. Neuron, 75, 265–270.

54. Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006). Predictive codes for forthcoming perception in the frontal cortex. Science, 314, 1311–1314.

55. Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. Nature Neuroscience, 19, 665–667.

56. Kok, P., & Turk-Browne, N. B. (2018). Associative prediction of visual shape in the hippocampus. Journal of Neuroscience, 38, 6888–6899.

57. Kumaran, D., & Maguire, E. A. (2006). An unexpected sequence of events: Mismatch detection in the human hippocampus. PLoS Biology, 4, e424.

58. Kumaran, D., & Maguire, E. A. (2007). Match–mismatch processes underlie human hippocampal responses to associative novelty. Journal of Neuroscience, 27, 8517–8524.

59. Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. Annals of the New York Academy of Sciences, 1124, 1–38.

60. Raichle, M. E. (2015). The brain's default mode network. Annual Review of Neuroscience, 38, 433–447.

61. Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., Bezgin, G., Eickhoff, S. B., Castellanos, F. X., Petrides, M., Jefferies, E., & Smallwood, J. (2016). Situating the

default-mode network along a principal gradient of macroscale cortical organization. Proceedings of the National Academy of Sciences of the United States of America, 113, 12574–12579.

62. Smallwood, J., Bernhardt, B. C., Leech, R., Bzdok, D., Jefferies, E., & Margulies, D. S. (2021). The default mode network in cognition: A topographical perspective. Nature Reviews Neuroscience, 22, 503–513.

63. Murray, J. D., Bernacchia, A., Freedman, D. J., Romo, R., Wallis, J. D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D., & Wang, X.-J. (2014). A hierarchy of intrinsic timescales across primate cortex. Nature Neuroscience, 17, 1661–1663.

64. Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: Memory as an integral component of information processing. Trends in Cognitive Sciences, 19, 304–313.

65. Yeshurun, Y., Nguyen, M., & Hasson, U. (2021). The default mode network: Where the idiosyncratic self meets the shared social world. Nature Reviews Neuroscience, 22, 181–192.

66. Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. Journal of Neurophysiology, 86, 1916–1936.

67. Kelly, S. P., & O'Connell, R. G. (2013). Internal and external influences on the rate of sensory evidence accumulation in the human brain. Journal of Neuroscience, 33, 19434–19441.

68. Hanks, T. D., & Summerfield, C. (2017). Perceptual decision making in rodents, monkeys, and humans. Neuron, 93, 15–31.

69. Siegel, M., Buschman, T. J., & Miller, E. K. (2015). Cortical information flow during flexible sensorimotor decisions. Science, 348, 1352–1355.

70. Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. Nature, 592, 601–605.

71. Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. Trends in Neurosciences, 23, 571–579.

72. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nature Neuroscience, 22, 974–983.

73. Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R., & Halgren, E. (2006). Top-down facilitation of visual object recognition. Proceedings of the National Academy of Sciences of the United States of America, 103, 449–454.

74. Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. Trends in Cognitive Sciences, 19, 394–405.

75. Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J.-D. (2017). The distributed nature of working memory. Trends in Cognitive Sciences, 21, 111–124.

76. Parvizi, J., & Kastner, S. (2018). Promises and limitations of human intracranial electroencephalography. Nature Neuroscience, 21, 474–483.

Chapter 3. Confidence Gates: A Re-examination of Evidence for False Feedback

1. Introduction

We rarely make a decision and leave it there. Having committed to a choice, we are routinely told whether it was right — by a colleague, an instrument, a grade, or, increasingly, an automated system — and we must then integrate that external verdict with a belief we already hold. How people do this has long been studied under the heading of *advice taking*, and the first-order picture is strikingly simple. People generally combine their own judgment with an outside opinion by averaging the two, taking a weighted blend of the estimates [1,2]. This strategy is robust because the average of two imperfect estimates tends to be at least as accurate as either estimate alone. Yet the blend is systematically lopsided: people anchor on their own view and move only partway toward the advice, a conservative under-use known as *egocentric discounting*, with the weight assigned to outside opinion typically falling well short of an even split [3]. In practice, this resembles a simple heuristic that nudges one's estimate a fixed fraction toward the verdict — an automatic, reliability-blind averaging process that reproduces the familiar under-weighting of advice regardless of whether the verdict deserves to be trusted.

But the verdicts we receive are not uniformly trustworthy. A colleague can be mistaken, a test can be miscalibrated, and an algorithm can err at rates that drift as conditions change [4].

Reliability is uncertain, often unsignaled, and rarely fixed. The challenge facing an efficient decision-maker is therefore not merely whether to incorporate external feedback, but how strongly to weight it given a reliability that must itself be inferred. When weighting is appropriate, combining two estimates can yield greater accuracy than relying on either alone [5,6]. Meeting this challenge requires some internal indication of how much weight one's own judgment deserves relative to the external verdict. Decision confidence offers precisely such a signal: by expressing how likely one believes one's judgment to be correct, it provides a basis for determining how readily conflicting feedback should be accepted.

A large body of work has shown that, after making a judgment, people can estimate how likely they are to have been correct, and that these confidence reports track performance across domains ranging from perceptual discrimination to memory, reasoning, and other higher-level cognitive tasks [7–9]. Although confidence is not perfectly calibrated, it provides a trial-by-trial metacognitive readout of the strength and reliability of one's own belief [10,11]. This makes it particularly well suited to feedback integration. If confidence is, as decision science proposes, an estimate of the probability that a choice is correct [12,13], then it is expressed in precisely the terms needed to compare one's own judgment with an external claim. Consistent with this view, confidence governs both how much an outside opinion is absorbed and how trust in a source develops across repeated encounters [14]. Moreover, when people exchange confidence rather than merely communicating their choices, confidence-weighted integration can approach statistical optimality [5]. Confidence may therefore serve not simply as a report about one's own decision, but as the internal signal through which the credibility of external feedback is evaluated and learned over time [15].

Although confidence appears to provide the signal from which the reliability of external feedback is inferred, what remains unclear is the mechanism on the far side of that inference: how belief revision is actually carried out once confidence has suggested that a verdict may be unreliable. A natural default account makes a strong but limiting prediction. If revision simply turns a "volume knob" on the verdict, the revised belief is computed from just two inputs: the reported confidence and the verdict. It follows that any two trials sharing the same confidence and the same verdict must end in the same revised belief. In particular, the revision cannot come out differently when the verdict happens to be false than when it is genuine — the revision process has nothing with which to tell the two cases apart, because the truth of the verdict is precisely what it never sees. Yet confidence is only a compressed and imperfect summary of the evidence underlying a decision [12,13]. This leaves open another possibility: rather than relying exclusively on this summary, the decision-maker may return to the evidence from which it was derived.

Several findings suggest that such a return is feasible. Evidence can remain available after a decision and continue to be processed: it may accumulate beyond the initial choice, producing changes of mind in the absence of new input [16,17]; a second, more deliberative confidence judgment may be better calibrated than an initial one [18,19]; and confidence itself may signal when further evaluation is worthwhile [15]. A verdict that conflicts with a confident judgment is therefore precisely the kind of event that could trigger a re-examination of the underlying evidence. This yields a prediction the volume-knob account cannot accommodate. Compare trials on which the reported confidence and the verdict are identical, but the verdict is genuine on some and false on others. Because the underlying evidence is richer than the single number into which it was compressed, a second look can tell these cases apart in a way the confidence report

alone cannot. Revised confidence should therefore end up systematically higher on trials where the initial choice was in fact correct — the revised belief tracks the truth even though everything the volume-knob account operates on is held fixed.

Taken together, these considerations suggest a confidence-gated process in which confidence serves a dual function in the exchange between internal and external information. In one direction, confidence is used to evaluate the external signal: the more confident one is in an initial judgment, the less credible a contradictory verdict becomes, because a verdict that opposes a near-certain judgment is more likely to be mistaken. In the other direction, the external signal acts on internal processing: a verdict that conflicts with a confident judgment serves as a cue to reopen the case and re-examine the evidence underlying the original belief. Confidence thus acts as a bidirectional coupling signal between internal and external information. It allows internal evidence to inform the evaluation of external feedback while also allowing external feedback to determine whether stored internal evidence should be revisited. We propose that this confidence-gated coupling enables people to detect and resist false feedback.

Our proposal unifies several ideas about feedback incorporation that have traditionally been studied in isolation. At the foundation lies heuristic averaging: a fixed, reliability-blind pull toward the verdict. Superimposed on this is confidence-based reliability estimation, together with the re-evaluation of post-decisional evidence when feedback calls the original judgment into question. Rather than treating heuristic averaging, confidence-based weighting, and post-decisional evidence processing as competing explanations, we propose that they operate jointly within a single feedback-integration mechanism.

We test this framework using a confidence–feedback–confidence (CFC) paradigm and a generative model designed around it. In the task, participants made a difficult perceptual

judgment, rated their confidence in that judgment (C1), received a verdict from an "automated grader" the reliability of which drifted covertly throughout the session, and then rated their confidence again (C2). Because the verdict was secretly inverted on a subset of trials, we know on every trial whether the feedback was genuine or false. This yields exactly the comparison the re-examination account's prediction calls for — that the revised belief should track the truth even with reported confidence and verdict held fixed: trials that share the same reported confidence and the same verdict but differ only in whether the verdict was true. If the second confidence judgment ends up systematically higher on the trials where the choice was in fact correct, it has recovered information about the true state of the world that the first report did not carry — and that recovery is precisely what it means to have partly detected the false feedback. This comparison provided an objective, trial-level measure of false-feedback detection. (Throughout the paper we use "detection" as shorthand for this quantity, false-feedback detection, and never for the perceptual detection of the stimulus.)

We formalized these processes in a single-observer generative model. The model first incorporates each verdict through a fixed-fraction heuristic averaging step. It then uses its own confidence to infer whether the verdict is likely to be genuine or inverted. When a contradiction is judged likely to reflect false feedback, the model re-examines its stored evidence and resists the verdict accordingly. False-feedback detection is therefore attributed specifically to the confidence-gated process. Critically, the model predicts rather than assumes that false-feedback detection should emerge primarily at high levels of confidence and predominantly following contradictory verdicts. The confidence threshold is determined by the grader's error rate rather than fitted as a free parameter, and the resulting predictions match the observed pattern of behavior.

Importantly, the same framework also extends naturally to individual differences in metacognitive efficiency—the fidelity with which confidence tracks accuracy [7,20]. Individuals differ substantially in how accurately their confidence distinguishes correct from incorrect judgments, and these differences have been associated with variation in behavior across multiple domains [21,22]. However, comparatively little is known about how metacognitive efficiency shapes the effectiveness of feedback integration. In our framework, confidence provides the signal that links internal evidence to the evaluation of an external verdict. The fidelity of that signal should therefore influence the coherence of feedback integration: individuals with higher metacognitive efficiency should be more likely to revise their beliefs in a consistent and appropriate direction.

The model thus seeks to explain not only how feedback is integrated within an individual, but also why some individuals integrate feedback more effectively than others. Consistent with this prediction, the same metacognitive-efficiency parameter governing both the initial and post-feedback confidence reports predicted the observed coherence of feedback integration across participants.

2. Methods

2.1 Participants

We recruited participants online through Prolific, a platform for paid behavioral research. All procedures were approved by the Institutional Review Board of the University of California, Los Angeles, and all participants gave informed consent before taking part. Eligibility was restricted to adults aged 18–40 with normal or corrected-to-normal vision, English as a first language, a

Prolific approval rating of 100%, and more than 50 prior approved submissions. Each person completed a single session lasting roughly an hour and received a fixed payment of $18 for taking part, plus a performance-based bonus of up to $6 that rewarded accurate choices and, especially, accurate confidence reports (the scoring rule is described at the end of Section 2.2). Paying for accurate confidence gives participants a concrete reason to report what they truly believe rather than responding carelessly.

Ninety people completed the full task. We then applied a set of quality-control criteria, fixed in advance, that were designed to remove participants who were not engaging with the task or whose data could not support the analyses (the criteria are detailed at the end of Section 2.2); 74 participants passed and form the basis of all analyses reported below.

2.2 Task and design

Each trial began with a brief display containing two patches of dots, and the participant decided which patch had more dots — a simple two-alternative perceptual judgment. Concretely, after a 500-ms fixation cross, two black squares appeared side by side for 300 ms, each containing a 25 × 25 grid of positions (625 in total), a subset of which was filled with white dots. The reference square was always half-filled (313 dots), while the other contained $313 + \Delta$ dots, with the more-numerous square equally often on the left or the right; participants then indicated, at their own pace, which side had contained more dots. We made this judgment deliberately difficult, and we tuned its difficulty continuously for each person using an adaptive one-up/two-down staircase: the task was made slightly harder only after two correct responses in a row and slightly easier after a single error, with equal step sizes up and down (in log space). Specifically, the staircase operated on $x = \ln(\Delta)$, starting at $x = 4.2$ ($\Delta \approx 67$ dots) and moving in steps of 0.4 log units over

the first six trials, 0.2 over the next five, and 0.1 thereafter, with Δ bounded between 1 and 70 dots. This rule converges on the difficulty at which the participant is correct about 71% of the time. The point of this procedure is to put everyone at roughly the same level of perceptual performance, so that differences between people reflect how they handle confidence and feedback rather than differences in their perceptual discrimination ability.

Immediately after choosing, the participant reported how confident they were that their choice was correct, on a sliding scale from 50% (a pure guess) to 100% (complete certainty). We call this first rating C1. The slider always began from a fresh random position on every rating, so participants could not simply leave it where it was, and the C1 value was never shown again.

Next, an "automated grader" delivered a verdict on the choice — either "CORRECT" or "INCORRECT". Participants had been introduced to this grader in the instructions (where it was labelled an "automated detector"): they were told that it would also count the dots and score each choice, and — verbatim — that it was "usually, but not always, accurate — it is an imperfect machine and sometimes scores a choice wrongly". Finally, the participant rated their confidence a second time, answering the very same question ("how confident are you that your choice was correct?") on the same 50–100% scale and again from a fresh, random slider start position; we call this second, post-feedback rating C2. Every trial is therefore a confidence → feedback → confidence sequence, and the question at the heart of the study is how the verdict in the middle moves a person from C1 to C2.

The main task consisted of 240 such trials (Figure 3-1). It was preceded by a 70-trial calibration block containing the same perceptual choice and a single confidence rating, but no grader and no feedback. The purpose of this block was to let the staircase settle on each person's difficulty level before the main task began.

A further manipulation concerned how trustworthy the grader actually was — and, importantly, this varied without the participant being told. Behind the scenes the session was divided into six hidden segments. In each segment the grader had a fixed "inversion rate", q: the probability that, on any given trial in that segment, it would flip its verdict and announce the opposite of the truth (calling a correct choice "incorrect", or vice versa). We used three levels of q — 0 (a perfectly reliable segment, never wrong), 0.3 (wrong on about a third of trials), and 0.5 (wrong half the time, i.e., uninformative). Each of the three levels was used in exactly two of the six segments, so that exposure to each level was matched at 80 trials over the session. The order of the six segments was a random permutation constrained so that no two adjacent segments shared the same reliability and the segment lengths were themselves randomized (each level's 80 trials were split between its two segments, with each part between 30 and 50 trials long). Note that this whole schedule — the order of the segments and their lengths — was generated afresh from a random seed for each participant, which decouples the grader's reliability from absolute position in the session and means no two people saw the same sequence of changes. Averaged over the whole session, the grader inverted its verdict on about 27% of trials ($\bar{q} = 0.267$). Participants were simply told that the grader was imperfect and that how much to rely on it was up to them; they were never shown the reliability schedule, and they never received any per-trial point feedback, because seeing their score on each trial would have revealed which verdicts had been inverted. Finally, at twelve unannounced points in the main task (two per hidden segment, never within the first five trials of a segment), participants were additionally asked, immediately after the C2 rating, to estimate the grader's current reliability — "Right now, how often do you think the detector is correct?" — on a slider from 0% (never) to 100% (always) that started at a

random position.

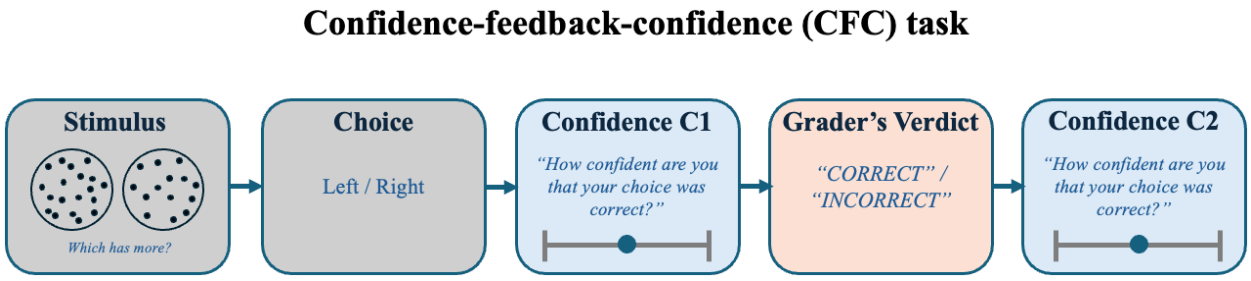

Figure 3-1. Confidence–feedback–confidence (CFC) task. Each trial: a brief dot-discrimination stimulus (two patches; the participant reports which contains more dots; the actual stimuli were two square boxes, drawn here as round patches for illustration), a left/right choice, a pre-feedback confidence rating (C1), a verdict from an "automated grader," and a post-feedback confidence rating (C2). The grader's reliability was covert and volatile: six hidden segments had true inversion rates $q \in \{0, 0.3, 0.5\}$, each level used in exactly two segments in a per-participant randomized order (experienced mean $\bar{q} = 0.267$), and on inverted trials the verdict was flipped relative to the truth.

Participants received the bonus payment based on their performance in the task which combined two kinds of points. Choice points added +1 for every choice that was actually correct, regardless of what the grader said. Confidence points rewarded accurate confidence through a standard proper scoring rule (the quadratic, or Brier, rule) [23]: each rating earned $1 - (\text{confidence}/100 - \text{outcome})^2$ points, where "outcome" is 1 if the choice was actually correct and 0 if not. Because a proper rule penalizes both over-stated and under-stated confidence, participants maximized their expected points only by reporting their true probability of being correct, and the rule was applied to both the C1 and the C2 rating, so well-judged confidence paid off at both moments.

We removed participants whose data could not support the analyses or who showed signs of disengagement: overall accuracy outside 0.60–0.85 (well below suggests disengagement; well above, a staircase that failed to make the task hard enough to produce informative errors);

calibration confidence ratings that were almost all the same value (a near-flat distribution, SD < 3.5, indicating the scale was not genuinely used); three or more failed attention checks; a staircase that never settled on a stable difficulty; or failure of a basic compliance check — in the perfectly reliable (q = 0) segments, where the grader is always right, any sensible participant should on average shift their confidence toward what the verdict says, and we excluded anyone who did not. After these participant-level exclusions, we also removed a small number of individual trials (1.1%) on which the confidence change was a wild outlier relative to that same participant's usual revisions — most likely lapses or mis-clicks — using a robust outlier rule (per-participant residuals beyond three standard deviations). This removed 1.1% of trials, leaving 17,564 trials from 74 participants for analysis.

2.3 Formal generative model

We built a computational model of a single observer and asked whether it can explain the observed behavior. We first sketch the model in words, then give its formal specification. On each trial the model sees the stimulus as one noisy sample of evidence, makes its choice, and reports a first confidence (C1) by reading out that evidence imperfectly. When the verdict arrives, the model revises its confidence in two distinct ways. Most of the revision is a cautious compromise: it shifts its confidence a fraction of the way toward what the verdict implies (toward certainty after "CORRECT", toward chance after "INCORRECT"); this is the average, or mean, revision. On top of that, it performs a check on whether the verdict can be trusted: it works out, from its own confidence and the grader's typical error rate, how likely it is that this particular verdict is a lie, and if it concludes the verdict is probably inverted it re-examines its memory of the stimulus and resists — and this is what produces false-feedback detection.

Finally, and importantly, the model reports its revised confidence (C2) through the very same noisy channel that produced C1, so the one quantity that blurs C1 also adds scatter to C2. Turning now to the formal specification: the observer is defined by the following generative process. A trial has a true state $d \in \{-1, +1\}$ (which side has more dots). The observer draws one noisy evidence sample and chooses the favored side:

$$x_1 \sim \text{Normal}(d \cdot v, 1); \quad a = \text{sign}(x_1); \quad \text{correct} = [a = d]$$

with drift v setting sensitivity ($d' = 2v$). The evidence carries a log-odds of being correct equal to $2v \cdot |x_1|$. This is turned into a reported confidence by a metacognitive readout that applies a gain g — over- or under-confidence, a property of how evidence is reported rather than of the decision itself, which uses only the sign of x_1 — together with readout noise σ:

$$C_1 = \text{sigmoid}(g \cdot 2v \cdot |x_1| + \text{Normal}(0, \sigma)), \text{ clipped to } [0.50, 0.995]$$

Here $g > 1$ yields over-confidence and σ sets metacognitive efficiency (M-ratio). The verdict is generated by inverting the true outcome with the segment's probability q, giving valence $F = +1$ ("CORRECT") or -1 ("INCORRECT"). Revision then has two levels, deliberately on two scales. Across participants, the mean of the estimated integration weights is a conservative linear pool toward the verdict-implied pole, on the probability scale:

$$C_2\_pool = (1 - w\_F) \cdot C_1 + w\_F \cdot anchor\_F$$

with a weight and anchor per valence (told-CORRECT moves toward certainty, told-INCORRECT toward chance). False-feedback detection operates on the latent log-odds scale.

First, the observer infers whether the verdict is genuine or inverted, using its own confidence and an assumed inversion rate $\hat{q}$:

$$P(\text{inverted} \mid C_1, F) = P(F \mid \text{inv}) \cdot \hat{q} \,/\, [\, P(F \mid \text{inv}) \cdot \hat{q} + P(F \mid \text{gen}) \cdot (1 - \hat{q}) \,]$$

$$\text{judge} = 1 \text{ if } P(\text{inverted} \mid C_1, F) > 0.5, \text{ else } 0$$

For a disconfirming verdict this crosses 0.5 at $C_1 = 1 - \hat{q}$; with $\hat{q} = 0.267$ the gate sits at ≈ 73%. When the observer concludes the verdict is inverted, it re-examines the evidence — re-reading the stored stimulus as a partially independent sample (correlation ρ with the original evidence):

$$x\_redel = d \cdot v + \rho \cdot (x_1 - d \cdot v) + \sqrt{(1 - \rho^2)} \cdot \varepsilon, \quad \varepsilon \sim \text{Normal}(0,1)$$

$$L\_redel = g \cdot 2v \cdot (a \cdot x\_redel)$$

$$\delta\_redel = L\_redel - E[L\_redel \mid C_1, F]$$

Here δ_redel is the surprise in the re-deliberation: how much the re-examined evidence deviates from what the observer's own confidence already implied, $E[L\_redel \mid C_1, F]$. False-feedback detection scales with this prediction error rather than with the raw re-read, which has two consequences. Mechanistically, the observer revises in proportion to how far a second look departs from its prior expectation — a within-trial quantity it has access to, not a comparison across other trials. Statistically, because the increment has zero expectation given C_1 and F, it is orthogonal to the mean revision by construction: the average revision is carried entirely by the linear pool, and δ_redel only redistributes confidence within a (valence × C1) cell according to whether the trial was, in fact, correct. In practice we approximate the conditional expectation $E[L\_redel \mid C_1, F]$ by the empirical mean within each (valence × C1-bin). Finally, C2 is reported

through the same noisy metacognitive channel as C1 (an independent draw of σ), which scatters the update and occasionally pushes it against the verdict:

$$C_2 = \text{sigmoid}(\ \text{logit}(C_2\_\text{pool}) + \kappa\_\text{redel} \cdot \text{judge} \cdot \delta\_\text{redel} + \text{Normal}(0, \sigma)\ )$$

The per-valence pool {w, anchor} is fit per participant by ordinary least squares of C2 on C1; the single false-feedback detection weight κ_redel is fit at the group level to the high-confidence disconfirming genuine/inverted gap. Three quantities are fixed rather than fitted: the drift v (set by the staircase), the assumed inversion rate $\hat{q}$ (set to the experienced average 0.267, so the gate is derived not fitted), and ρ (validated by a robustness sweep showing the observed false-feedback detection is reached at $\rho \approx 0.5$). Representative values were $v = 0.553$ ($d' \approx 1.1$), $g = 1.85$, $\sigma = 0.45$, $w \approx 0.29$–0.37, $\kappa\_\text{redel} = 0.058$. Crucially, σ is a single parameter shared by C1 (setting M-ratio) and C2 (setting integration coherence); no separate parameter couples metacognition to coherence. One further modelling choice is worth making explicit: the genuineness inference uses the reported C_1 at face value — the observer does not discount its own metacognitive noise σ when judging whether a verdict is inverted. Consequently σ leaves the magnitude of false-feedback detection essentially unchanged and acts almost entirely on coherence. This is deliberate: an idealized observer that discounted C_1 by σ would detect less at higher noise, implying a positive M-ratio–false-feedback detection relationship — which the data do not show (Section 3.4). Taking C_1 at face value is what keeps false-feedback detection σ-invariant, matching that null.

A note on $\hat{q}$: because participants are never told the grader's error rate, we interpret it not as knowledge the observer is given, nor as a value learned online over the session, but as an assumed prior the observer brings to the task — most plausibly anchored by the instruction that

the grader is imperfect, together with generic skepticism. Consistent with this, false-feedback detection is present from the earliest trials (Section 4.5) rather than ramping up over the session as an online-learned estimate would predict; and, as the reliability-invariance result shows, this assumed rate is then held roughly fixed rather than being re-estimated block by block.

3. Results

3.1 Mean revision: conservative and underweighted

By design, the staircasing held everyone's choice accuracy near 71%. Participants were nonetheless markedly over-confident: mean pre-feedback confidence (C1) was 80.8%, well above accuracy, with substantial variation across people (between-subject SD ≈ 10 points); mean post-feedback confidence (C2) was similar, at 82.3%. Because the grader erred often, participants were told "INCORRECT" on 40% of trials, giving a large supply of disconfirming verdicts.

The verdict moved confidence in the expected direction — up after "CORRECT", down after "INCORRECT" — but by much less than it should have. To quantify this effect, we compared each person's actual confidence changes with those of a learning-free ideal integrator, a benchmark observer that starts from the participant's own stated confidence and revises it by Bayes' rule given the grader's session-average reliability (73% correct). For example, if a participant reported 80% confidence and was then told "INCORRECT", the benchmark would lower confidence to about 59%; a participant who instead stopped at about 71% made roughly half of the prescribed revision (all changes are computed in log-odds units). The integration weight η summarizes this across a person's trials as the slope of actual on prescribed changes —

$\eta = 1$ means revising exactly as much as the benchmark, $\eta < 1$ under-using the feedback. The mean was $\eta = 0.51$ (median 0.46). In plain terms, people revised in the right direction, but only about half as far as the feedback warranted. This conservative under-use of outside information is the "egocentric discounting" long documented in advice taking [1–3], and it behaves here like a default assimilation heuristic — a cheap, always-on partial acceptance of the verdict that runs whether or not the verdict deserves trust. Notably, this conservatism barely changed as the grader's reliability drifted: the mean verdict-aligned confidence change was 0.92, 0.88 and 0.87 log-odds units in blocks with $q = 0$, 0.3 and 0.5, respectively — far less variation than a participant applying the current reliability would show.

3.2 False-feedback detection: high-confidence and conflict-specific

Beyond the average pull toward the verdict, we next asked whether participants could detect false — that is, inverted — feedback, and whether such false-feedback detection is visible in the revised confidence. The logic of the test is as follows. Take all trials on which people reported the same C1 and received the same verdict: on some of these trials the verdict was genuine, on others it was a lie. From the participant's point of view the two sets are indistinguishable at the moment of revision — everything they could see, their own stated confidence and the verdict, is identical — and only the hidden truth of the verdict differs. Any revision strategy that works only from C1 and the verdict must therefore produce, on average, the same C2 in both sets. If C2 instead ends up systematically *higher* on the trials where the choice was *actually correct*, the revision must have drawn on information beyond those two inputs — the participant has, at least in part, seen through the false verdict. We therefore measure false-feedback detection as the difference in C2 between actually-correct and actually-wrong trials at matched C1 and verdict

type: within narrow 5-percentage-point bins of C1, we take the mean C2 of actually-correct trials minus that of actually-wrong trials, and average the differences across bins.

False-feedback detection was present, but only under specific conditions (Figure 3-2, Table 3-1). For disconfirming "INCORRECT" verdicts delivered at low confidence (C1 < 75%), there was no sign of it: matched on confidence, actually-correct trials ended no higher than actually-wrong ones (−0.29 points, p = .53). At high confidence (C1 ≥ 75%) it clearly emerged: actually-correct trials held +1.35 points higher (95% CI [0.57, 2.14], p = .001). A complementary regression confirmed it: predicting actual accuracy from standardized C1 and C2 together, the coefficient on C2 — its information about accuracy over and above C1 — was reliably positive on these trials (+0.06, 95% CI [0.04, 0.09], p < .001). Confident participants, in other words, partly saw through a false "INCORRECT". For confirming "CORRECT" verdicts, by contrast, false-feedback detection was negligible even at high confidence (+0.21, 95% CI [−0.10, 0.52], p = .20) — about a sixth of the disconfirming effect.

Cell (matched on C1)	False-feedback detection (pts)	95% CI	p
Told INCORRECT, C1 < 75	−0.29	[−1.19, 0.61]	.53
Told INCORRECT, C1 ≥ 75	+1.35	[0.57, 2.14]	.001
Told CORRECT, C1 < 75	+0.30	[−0.63, 1.23]	.53
Told CORRECT, C1 ≥ 75	+0.21	[−0.10, 0.52]	.20

Table 3-1. Matched false-feedback detection (subject-level mean C2 of actually-correct minus actually-wrong trials at equal C1), by verdict valence and confidence. False-feedback detection appears only at high confidence and overwhelmingly for disconfirming verdicts; the told-CORRECT effect is small and not significant.

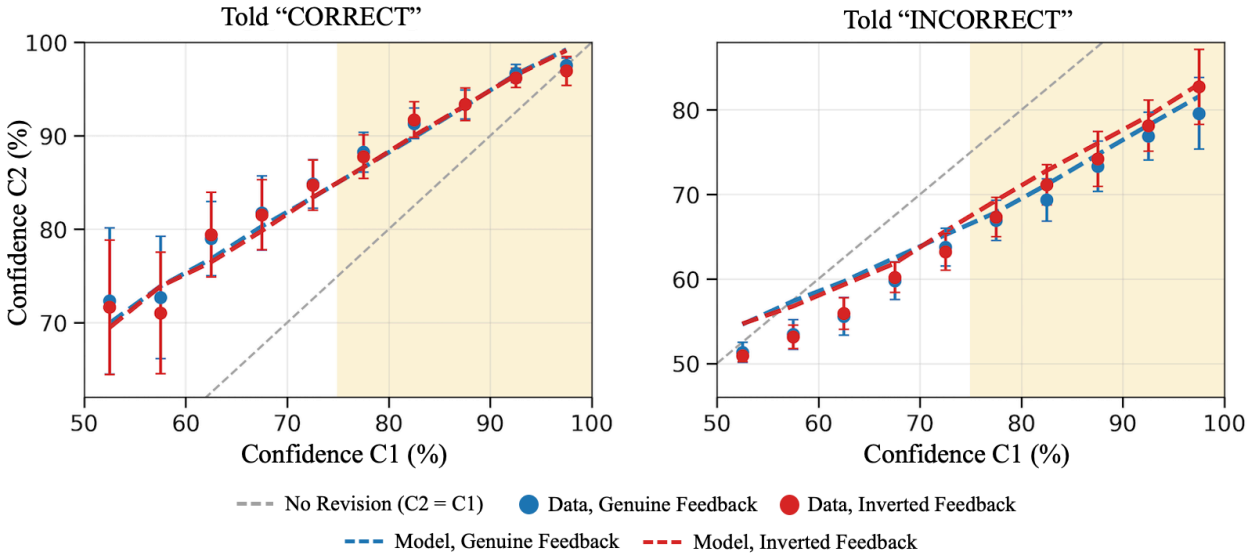

Figure 3-2. Confidence revision by verdict valence and the hidden truth of the feedback: data (points, 95% CI) versus the single-observer generative model (dashed lines, fixed theory-derived parameters simulated forward on the real schedules). Post-feedback confidence (C2) is plotted against pre-feedback confidence bin (C1), separately for genuine (blue) and inverted (red) feedback. Points and lines are subject-averaged bin means, and within each bin both series are restricted to the same participants (or simulated observers) — those contributing at least two trials of each type — so that the vertical gap between them is a within-participant contrast, as in Table 3-1. Error bars are 95% CIs on each series, not on the gap. The two panels use different y-axis ranges. Shading marks the high-confidence region (C1 ≥ 75%). In the told-INCORRECT panel, inverted feedback (where the participant was actually correct) holds C2 above genuine feedback at high confidence — partial resistance to false verdicts, reproduced by the model. The told-CORRECT divergence is much smaller.

The comparisons so far are descriptive, taking one confidence-by-valence cell at a time. Our primary inferential test therefore estimates false-feedback detection and its moderators jointly, in a single trial-level model of C2 on high-confidence trials in the partly-unreliable blocks ($q \in \{0.3, 0.5\}$; the fully reliable blocks are excluded, since with no inversions the verdict and the truth are equivalent). It controls for C1 and lets false-feedback detection interact with verdict valence and with the current and recently experienced (previous 30 trials) inversion rates, and was fit within each participant and tested across the sample; a pooled model with cluster-robust standard errors agreed (Table 3-2). False-feedback detection was significant ($\beta = +0.99$, $p =$

.002) and far larger for disconfirming than confirming verdicts (detection × valence, β = +2.01, p < .001); estimated within each valence, the simple effect was strong for disconfirming verdicts (+2.00, p < .001) and essentially zero for confirming ones (−0.01, p = .97), in agreement with the confidence-matched comparison above. (We quote Table 3-1's confidence-matched values as the false-feedback detection magnitudes, because the regression is fit on raw C2, where told-CORRECT sits near the response ceiling.) By contrast, neither reliability term moderated false-feedback detection: it was no stronger when the current block was less reliable (detection × block, β = +0.99, p = .14) and did not scale with the recently experienced inversion rate (β = −0.29, p = .49; Table 3-2). This insensitivity is notable because participants were not oblivious to the reliability changes: their answers to the periodic probe question tracked the grader's current reliability (mean within-subject r = 0.33, p < .001, positive in 85% of participants). Explicit monitoring and trial-by-trial revision thus appear to come apart — a contrast we take up in the Discussion (Section 4.4).

Mixed-model term	β	t	p
False-feedback detection (actually correct)	+0.99	3.19	.002
× verdict valence (told-INCORRECT)	+2.01	4.55	<.001
× current block (q0.5 vs q0.3)	+0.99	1.51	.14
× recent inversion rate (per SD)	−0.29	−0.70	.49

Table 3-2. Key effects from the trial-level analysis (7,847 high-confidence trials from 74 participants in q ∈ {0.3, 0.5}; C2 as outcome). Coefficients are the mean of per-subject estimates, tested across participants (t); a pooled subject-fixed-effects model with cluster-robust SEs agrees. False-feedback detection and its conflict-specificity are significant; neither the current block's reliability nor the recently-experienced inversion rate significantly modulates false-feedback detection.

3.3 Single-observer model fitting and comparison

Why does false-feedback detection happen only at high confidence, and only for contradictions? Our single-observer generative model proposes that false-feedback detection is a small inference

the observer runs on each trial: given how confident I was, and how often this grader errs, how likely is it that this particular verdict is a lie? The intuition needs no mathematics. If you were 95% sure and the grader says “INCORRECT”, either you genuinely erred or the grader inverted — and your own confidence makes the second reading the likelier one, so you hold your ground. If you were only 55% sure, the same verdict is perfectly believable, and you accept it.

Worked through formally, a disconfirming verdict becomes more likely a lie than the truth exactly when confidence exceeds $1 - \hat{q}$, where $\hat{q}$ is the grader’s assumed error rate. Plugging in the error rate participants actually experienced ($\hat{q} = 0.267$) puts the threshold at about 73% confidence — essentially where false-feedback detection switched on in the data. Nothing was fitted to produce this number: the location of the high-confidence onset is a prediction inherited from the grader’s error rate. When the observer concludes a verdict is probably inverted, it takes a second look at its stored impression of the stimulus and uses that to resist.

The same inference explains the valence asymmetry. To suspect a “CORRECT” verdict of being a lie, you would have to believe you were probably wrong despite having just chosen — confidence below about 27% — which the scale cannot even express, since confidence in one’s own choice cannot fall below 50%. Confirming verdicts are therefore essentially never flagged, and conflict-specificity follows with no extra assumption. Run forward on each participant’s own reliability schedule, the model reproduces the data closely (Figure 3-2): the 71% accuracy, the mean confidence levels (C1 ≈ 80, C2 ≈ 81), the shape of both revision curves, high-confidence disconfirming false-feedback detection of +1.36 points (data: +1.35), and near-zero false-feedback detection in the other three cells — including the confirming-verdict cell, where the model’s prediction of essentially nothing is borne out (model ≈ 0; data +0.21, not significant).

Reproducing the data, however, does not by itself show which parts of the model are doing the work. We therefore tested its two essential ingredients — the confidence gate, and the re-examination of stored stimulus evidence that it triggers — by lesioning each in turn and asking how false-feedback detection changes. We compared the full model against three variants built on identical machinery, each given the same false-feedback detection weight where one applies, and scored all four on the four false-feedback detection cells (Figure 3-3): a trust/weighting model with no re-examination, in which C2 is only a confidence-weighted compromise between C1 and the verdict — the standard advice-taking account, in which confidence does nothing but set how far to move; a re-examination that re-reads the same evidence without drawing a fresh sample ($\rho = 1$); and an ungated re-examination that runs on every trial, conflict or not.

Each variant fails informatively. The trust/weighting model produces essentially no false-feedback detection (high-confidence disconfirming +0.14, versus +1.35 in the data): any rule that maps C1 and the verdict onto C2 leaves nothing at matched C1 that can track the truth, so confidence-as-trust alone cannot generate false-feedback detection. Re-reading the same evidence is nearly as weak (+0.28): the re-examination must draw partially independent evidence from stimulus memory rather than re-use what the confidence report already carried. The ungated variant produces false-feedback detection everywhere — including at low confidence and for confirming verdicts (told-CORRECT, low confidence: +1.29) — which the data contradict. Only the full model reproduces the observed localization (four-cell RMSE 0.16, versus 0.55–0.77 for the alternatives). False-feedback detection therefore needs both ingredients: a re-examination of the evidence itself, not a re-weighting of confidence, and a gate that engages only when feedback conflicts with a confident judgment.

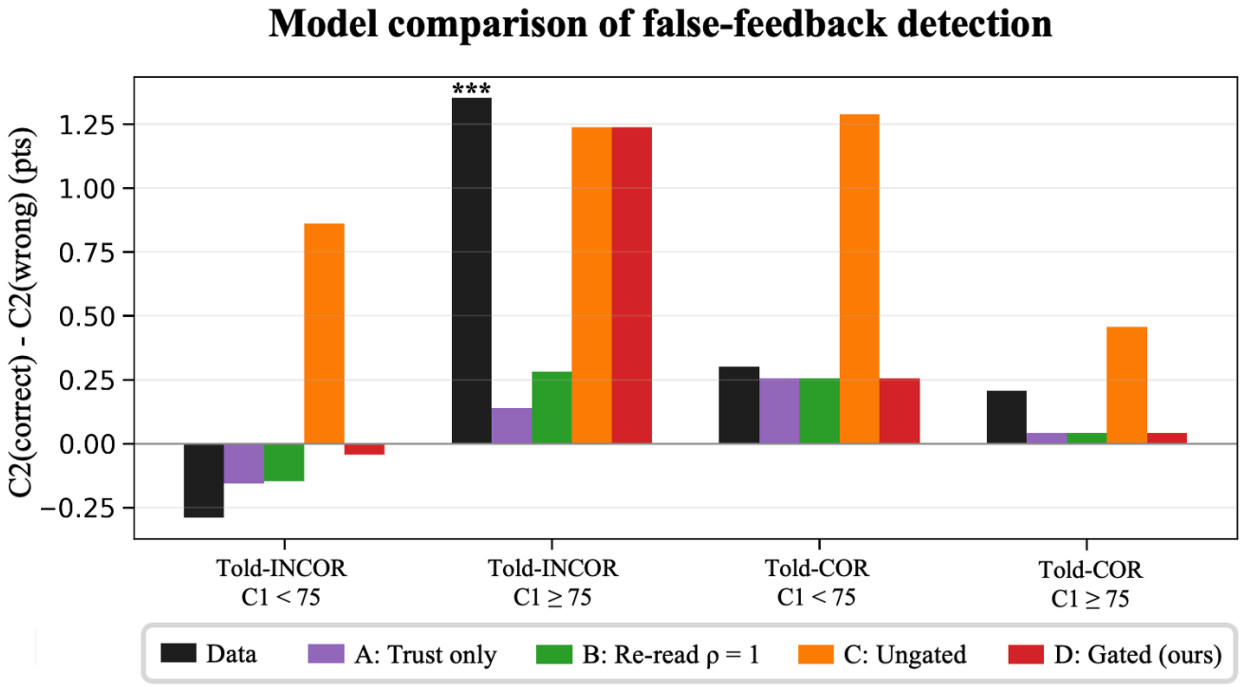

Figure 3-3. Model comparison. False-feedback detection (matched C2 of actually-correct minus actually-wrong trials) in the four confidence × valence cells, for the data (per-subject matched, as in Table 3-1) and four models that share identical machinery: A, trust/weighting only (no re-examination); B, re-examination that re-reads the same evidence (ρ = 1); C, ungated re-examination (every trial); and D, the full model — a confidence-gated re-examination of partially-independent stimulus evidence. Only D reproduces the data: false-feedback detection localized to high-confidence disconfirming verdicts. A and B produce almost no false-feedback detection; C produces detection in every cell.

3.4 Metacognitive efficiency predicts integration coherence

We next asked whether metacognitive efficiency — how faithfully a person's confidence tracks their own accuracy — shapes how participants handle feedback. We quantified it with the standard signal-detection measure meta-d′, expressed as the M-ratio (meta-d′/d′): an M-ratio of 1 means confidence is as informative about accuracy as an ideal observer's would be, and lower values mean it tracks accuracy more weakly [7,20]. Each person's M-ratio was computed from all 310 pre-feedback ratings they gave — the 70 calibration ratings plus the 240 C1 ratings from

the main task. Because every rating entering the measure precedes the verdict on its trial, M-ratio is uncontaminated by the feedback manipulation.

Metacognitive efficiency did not predict false-feedback detection (we return to why in the Discussion). What it predicted — strongly — was the coherence of revision: whether updates went in a sensible direction at all. Some revisions run against the verdict — confidence rising after "INCORRECT", or falling after "CORRECT" — and the fraction of such wrong-way moves is our measure of integration coherence, capturing not how much people moved but whether they moved sensibly. M-ratio correlated −0.40 with that fraction ($p < .001$; Figure 3-4, left), and −0.42 with the magnitude-weighted version, which additionally weights each wrong-way move by its size ($p < .001$; Figure 3-4, right). Concretely, the least metacognitively efficient third of participants moved the wrong way on roughly 11% of trials, the most efficient third on only about 6%. The relationship survived controlling for differences in mean confidence, in confidence variability, and in how often ratings changed at all (partial r = −0.40, $p < .001$).

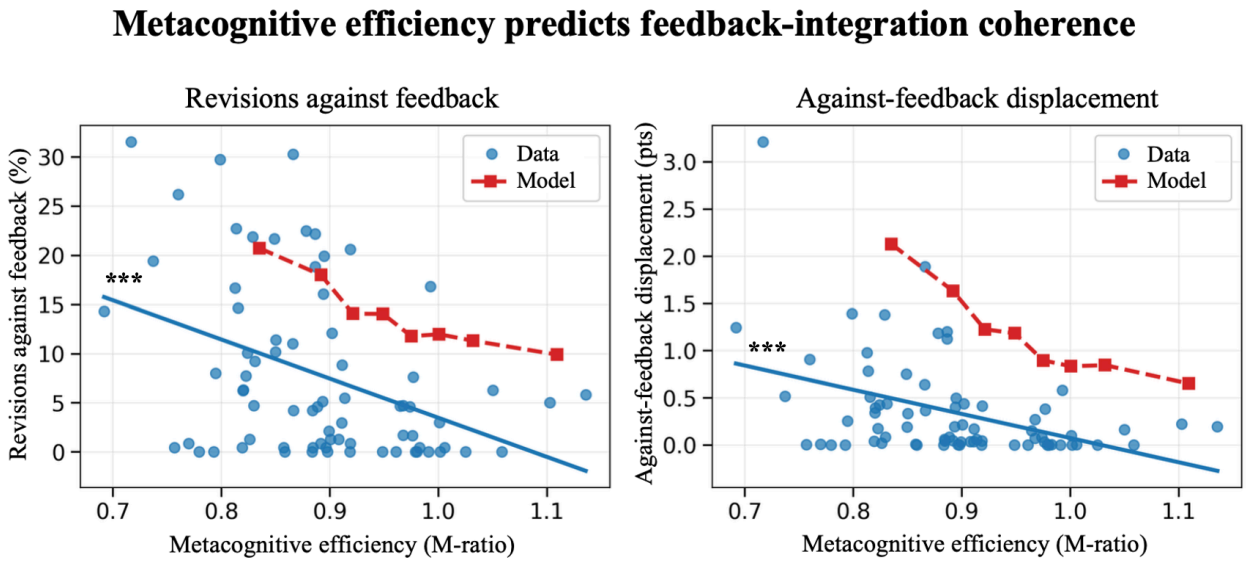

Figure 3-4. Metacognitive efficiency (M-ratio) predicts feedback-integration coherence. Left: fraction of revisions against the feedback; right: against-feedback displacement (frequency × magnitude). Blue points and

line: data. Red dashed line: the single-observer generative model with C2 readout noise σ_rep varied across simulated participants. Both data and model show fewer incoherent revisions at higher M-ratio.

The single-observer generative model produces this individual difference without any added machinery. Both confidence reports pass through the same noisy readout channel, and the level of that noise is exactly what sets a person's M-ratio. One quantity therefore does double duty: more readout noise makes confidence track accuracy more weakly, and it makes the post-feedback report more likely to be swamped by noise and land on the wrong side of the verdict. Varying only this noise level across simulated participants reproduced both the realistic rate of wrong-way revisions (about 10% at typical noise levels; 7.8% in the data) and the negative M-ratio–incoherence relationship ($r = -0.36$ to -0.37, matching the estimate in the data; Figure 3-4).

The upshot is a clean dissociation. Metacognitive ability does not make people better at detecting inaccurate feedback — the model treats that computation as available to everyone — but it governs how faithfully the result survives being turned into a report.

4. Discussion

Can people detect inaccurate feedback using nothing but their own confidence? Yes, but only within limits that are themselves informative. When the grader contradicted a confident judgment, the revised confidence tracked the truth beyond what the first report had already conveyed: at matched C1 and verdict, C2 was reliably higher when the choice had in fact been correct. False-feedback detection appeared only above roughly 75% confidence and almost only for disconfirming verdicts, and we predicted both boundaries rather than fitting them. The model

comparison showed that false-feedback detection works by re-consulting stored stimulus evidence: when confidence only set how much weight the verdict received, no false-feedback detection appeared at matched confidence. Metacognitive efficiency, by contrast, did not affect false-feedback detection, but it strongly predicted how coherently feedback was integrated. Below we relate these findings to earlier work on decision confidence.

4.1 Confidence as the arbiter between internal and external evidence

The average behavior in our task reproduces the most robust finding in the advice-taking literature: participants took the verdict on board by a fixed, conservative fraction, the egocentric discounting reported across decades of judge–advisor studies [1–3]. What the covert-inversion design adds is a view of the selective process that operates on top of this default. Because inverted verdicts gave us a trial-by-trial ground truth, we could show that people do more than merely down-weight a source they distrust on average [14]: they assess individual verdicts, and they do so in the currency that formal accounts say confidence provides — an estimate of the probability that the choice was correct [11–13,24,25]. Read as such a probability, confidence supports a simple inference: a contradiction from a source that errs at rate $\hat{q}$ is more likely false than true whenever confidence exceeds $1 - \hat{q}$. Two features of the data follow from this with no free parameters. False-feedback detection should begin near 73% confidence, and it should be almost absent for agreeing verdicts, because suspecting one would require confidence below 27% — a value the scale cannot express. That behavior matched both boundaries is, as far as we know, the most direct evidence to date that decision confidence is used online, verdict by verdict, to weigh internal evidence against an external claim [14].

4.2 A confidence-gated return to the evidence

How is the false-feedback resistance carried out? The model comparison points away from relying on the report-level arithmetic in the confidence rating and toward returning to the stimulus evidence itself. A variant in which confidence merely set the weight given to the verdict produced no false-feedback detection — any rule mapping C1 and the verdict onto C2 leaves nothing at matched C1 that can track the hidden truth — and a variant that re-read the same evidence without a fresh sample barely improved on it. False-feedback detection required a partially independent re-sample of stimulus information, engaged selectively by conflict. This dovetails with accumulating evidence that the information behind a decision remains available and continues to be processed after commitment: accumulation continues past the choice, producing changes of mind without new input [16,17,26]; these post-decisional revisions have identifiable neural mediators [27]; and a second, more deliberate confidence judgment is better calibrated than a first [18,19]. Our results suggest one function of this post-decisional machinery: an external contradiction can recruit it to audit the contradicting message itself.

This raises the question of what substrate holds the evidence that is re-read, and the timing of our task makes the answer fairly specific. The stimulus was present for only 300 ms, the verdict did not arrive until roughly 3 s after its offset, and the second rating was submitted about 7 s after it (median across participants) — far beyond the lifetime of iconic sensory memory. Whatever is consulted when a verdict conflicts with a confident judgment must therefore be a durable but capacity-limited representation, which is to say visual working memory [35]. Contemporary accounts make working memory a plausible substrate for exactly the operation our model requires. Its contents are not held as an abstract summary but maintained in distributed sensory and frontoparietal codes that remain decodable across a delay [36,37], and their fidelity is graded

rather than all-or-none: memory strength varies continuously from trial to trial, and each read-out of a stored trace is corrupted by noise [38,39]. A second consultation of such a trace yields precisely what the model assumes — a sample correlated with, but not identical to, the evidence behind the first confidence report ($\rho \approx 0.5$ in our simulations) — rather than either a fresh independent observation or a redundant copy of what confidence has already expressed. The re-examination is, on this reading, a second draw from an internally maintained distribution, close in spirit to the sampling accounts developed to explain why averaging a person's own repeated estimates improves them [18,40]. The model further requires the observer to appraise what that second look returns, judging whether the re-read evidence departs from what its own confidence implied, and people demonstrably have such access: introspective judgments predict the precision of individual working memory representations [41], and trial-to-trial memory uncertainty is carried forward into subsequent decisions in near-optimal fashion [42].

Working memory is also the natural arena for the encounter between maintained internal content and incoming external information. Gating accounts of prefrontal–striatal function describe this trade-off directly: maintained representations must be shielded from interference yet updated selectively when a task-relevant event warrants it, with separable input and output gates governing what enters the store and what is allowed to drive behavior [43,44]; related work treats the removal and replacement of outdated content as an active operation rather than passive decay [45]. The confidence-gated check we describe has this character. An arriving verdict does not simply overwrite the belief it addresses; a conflict with a confident judgment is the event that opens access to the stored evidence, and what that evidence returns then determines how far the belief moves. Individual differences point the same way: working memory capacity predicts

resistance to misinformation in the continued-influence paradigm [46], suggesting that how firmly information is held shapes how far an external claim can rewrite it.

This reading also yields predictions our design cannot test, since we did not measure working memory. A concurrent load imposed between the choice and the verdict should selectively reduce false-feedback detection while leaving the average assimilation intact, because the linear pool needs only C1 and the verdict, not the trace; lengthening the interval before the verdict should erode false-feedback detection as the representation degrades; and individual differences in working memory precision should predict the magnitude of false-feedback detection — where metacognitive efficiency, as we saw, did not. Confirming that profile would supply converging evidence for an evidence-level operation that our model comparison currently supports on behavioral grounds alone.

The gating we observe offers an instructive contrast with confirmation bias. High confidence has been shown to amplify the neural processing of choice-consistent evidence and mute disconfirming evidence — a confirmation bias that entrenches beliefs against valid opposition [28]. In our task the same signal played a protective role: high confidence licensed skepticism toward feedback that was, in fact, more likely to be false. The two findings may reflect one mechanism operating in different environments — where disconfirming signals are usually valid, confidence-gated discounting is a bias; where they are often lies, it is accuracy-preserving inference. More broadly, confidence is known to govern whether further information is sought at all [15,29] and can act as an internal teaching signal when external feedback is absent [30]; the present results add the complementary role of quality controller when external feedback is present but of uncertain provenance.

4.3 Metacognitive efficiency: computation versus fidelity

Metacognitive efficiency (M-ratio [7,20]) did not predict false-feedback detection, which at first seems paradoxical — false-feedback detection is a metacognitive feat, and the metacognitively gifted might be expected to excel at it. The model explains why this null is the expected result. As in recent accounts that locate metacognitive inefficiency in noise added at the readout of decision evidence [10,31,32], a single noise source both limits how well confidence tracks accuracy and scatters the post-feedback report. Because the genuineness inference takes the reported confidence at face value, readout noise leaves the average size of false-feedback detection intact; what it corrupts is the coherence of the reported revision, occasionally sending it against the verdict altogether. Exactly this dissociation appeared: M-ratio was unrelated to false-feedback detection but correlated −0.40 with the frequency of wrong-way revisions, a relationship that controls for confidence level and confidence variability. Individual differences in metacognitive efficiency, which vary widely across people and have been linked to brain structure [9,21,22], thus acquire a specific functional consequence for feedback integration: they determine not whether the confidence-based computation runs, but how faithfully its output survives being turned into a report. In applied terms, low metacognitive efficiency should manifest less as gullibility toward false feedback than as erratic, direction-inconsistent updating.

4.4 Monitoring reliability without using it

The grader's reliability drifted across six hidden segments, yet no analysis found revision behavior tracking it: verdict-aligned updating declined only marginally across reliability levels, and false-feedback detection was modulated neither by the current block nor by the recently experienced inversion rate. Strikingly, this was not ignorance: explicit probe estimates tracked

the grader's current reliability (mean within-subject r = 0.33), so participants monitored the drift even as their trial-by-trial weighting ignored it. People demonstrably can learn the reliability of information sources in volatile environments when the task centers on that learning [33,34], and confidence-based trust formation across repeated encounters is well documented [14]. Our results mark a boundary condition: when each trial supplies its own confidence-based check, the within-trial computation appears to satisfice, and explicitly monitored knowledge goes unused. Incentives likely matter here — given participants' conservative weights, fully exploiting the current reliability would have improved outcomes only marginally, so the invariance may reflect rational inattention rather than incapacity.

4.5 Limitations and future directions

Several limits bound these conclusions. False-feedback detection, though statistically robust, was modest in absolute size (about 1.4 points on a 50-point confidence scale) and comes from a single perceptual task with one reliability schedule; richer materials, and social rather than mechanical sources, are needed to establish generality. False-feedback detection was also strongest early in the session and waned thereafter, a decline whose source — fatigue or strategic disengagement — we cannot yet identify. Finally, whether people can be moved from monitoring reliability to using it — through stronger incentives, or prompts that link their explicit estimates to the next verdict — is an open question with practical stakes.

References

1. Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35(3), 780–805.
2. Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. Organizational Behavior and Human Decision Processes, 101(2), 127–151.
3. Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. Organizational Behavior and Human Decision Processes, 93(1), 1–13.
4. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80.
5. Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. Science, 329(5995), 1081–1085.
6. Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415(6870), 429–433.
7. Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. Frontiers in Human Neuroscience, 8, 443.

8. Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), The psychology of learning and motivation (Vol. 26, pp. 125–173). Academic Press.

9. Mazancieux, A., Fleming, S. M., Souchay, C., & Moulin, C. J. A. (2020). Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks. Journal of Experimental Psychology: General, 149(9), 1788–1799.

10. Shekhar, M., & Rahnev, D. (2021). The nature of metacognitive inefficiency in perceptual decision making. Psychological Review, 128(1), 45–70.

11. Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. Nature Neuroscience, 19(3), 366–374.

12. 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.

13. Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. Psychological Review, 124(1), 91–114.

14. Pescetelli, N., & Yeung, N. (2021). The role of decision confidence in advice-taking and trust formation. Journal of Experimental Psychology: General, 150(3), 507–526.

15. Desender, K., Donner, T. H., & Verguts, T. (2021). Dynamic expressions of confidence within an evidence accumulation framework. Cognition, 207, 104522.

16. Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. Psychological Review, 117(3), 864–901.

17. Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. Nature, 461(7261), 263–266.

18. Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. Psychological Science, 19(7), 645–647.

19. Elosegi, P., Rahnev, D., & Soto, D. (2024). Think twice: Re-assessing confidence improves visual metacognition. Attention, Perception, & Psychophysics, 86(2), 373–380.

20. Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Consciousness and Cognition, 21(1), 422–430.

21. 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.

22. Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human metacognition across domains: Insights from individual differences and neuroimaging. Personality Neuroscience, 1, e17.

23. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3.

24. Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. Neuron, 88(1), 78–92.

25. Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. Neuron, 90(3), 499–506.

26. van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. eLife, 5, e12192.

27. Fleming, S. M., van der Putten, E. J., & Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. Nature Neuroscience, 21(4), 617–624.

28. Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. Nature Communications, 11, 2634.

29. Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. Psychological Science, 29(5), 761–778.

30. Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. eLife, 5, e13388.

31. Guggenmos, M. (2022). Reverse engineering of metacognition. eLife, 11, e75420.

32. Boundy-Singer, Z. M., Ziemba, C. M., & Goris, R. L. T. (2023). Confidence reflects a noisy decision reliability estimate. Nature Human Behaviour, 7(1), 142–154.

33. Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. Nature Neuroscience, 10(9), 1214–1221.

34. Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. Nature, 456(7219), 245–249.

35. Baddeley, A. (2012). Working memory: Theories, models, and controversies. Annual Review of Psychology, 63, 1–29.

36. Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J.-D. (2017). The distributed nature of working memory. Trends in Cognitive Sciences, 21(2), 111–124.

37. Rademaker, R. L., Chunharas, C., & Serences, J. T. (2019). Coexisting representations of sensory and mnemonic information in human visual cortex. Nature Neuroscience, 22(8), 1336–1344.

38. Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. Nature Neuroscience, 17(3), 347–356.

39. Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. Nature Human Behaviour, 4(11), 1156–1172.

40. Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. Cognitive Science, 38(4), 599–637.

41. Rademaker, R. L., Tredway, C. H., & Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. Journal of Vision, 12(13), 21.

42. Honig, M., Ma, W. J., & Fougnie, D. (2020). Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions. Proceedings of the National Academy of Sciences, 117(15), 8391–8397.

43. O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. Neural Computation, 18(2), 283–328.

44. Chatham, C. H., & Badre, D. (2015). Multiple gates to working memory. Current Opinion in Behavioral Sciences, 1, 23–31.

45. Oberauer, K. (2019). Working memory and attention — A conceptual analysis and review. Journal of Cognition, 2(1), 36.

46. Brydges, C. R., Gignac, G. E., & Ecker, U. K. H. (2018). Working memory capacity, short-term memory capacity, and the continued influence effect: A latent-variable analysis. Intelligence, 69, 117–122.

Chapter 4. "I Must Have Wanted That": One's Own Action as Evidence in Metacognitive Goal Inference during Value-Based Decision Making

1. Introduction

Metacognition — the capacity to monitor and evaluate one's own cognitive processes [1, 2] — has become an increasingly important component of research on value-based decision making. Most work has focused on decision confidence: how people evaluate the reliability of their own preference-based choices, how such confidence reflects the strength of the underlying value evidence, and how it guides subsequent behavior [3–5]. This work has established that value-based decisions are accompanied not only by first-order computations about what to choose, but also by monitoring of those computations. Yet decision confidence concerns only one property of a decision — whether it is likely to be correct or reliable — a restriction increasingly noted within metacognition research itself [6, 7]. Value-guided behavior contains other latent states that an agent may also need to infer about itself.

One particularly important candidate is the goal of an action itself. Value-based behavior is organized around learned values that make some outcomes more desirable than others [8, 9], but the goal underlying a particular action is not necessarily transparent even to the person who performed it. Intentions may be weak, mixed, or fleeting; actions are corrupted by motor noise;

and their consequences can diverge from what was originally intended. Classic evidence confirms that this opacity is real rather than hypothetical: people confabulate reasons for behavior shaped by factors they cannot report [10], and when choices are covertly swapped, they fluently justify decisions they never made [11, 12]. Determining "what was I trying to obtain?" therefore requires a judgment about the content of one's own internal state. In this sense, a goal report can be understood as a metacognitive judgment just as decision confidence can, differing only in what the inference concerns: not a property of a first-order state — its likely accuracy — but its content. Such judgments may be especially consequential in value-based behavior, because what people believe they wanted can influence subsequent preferences, choices, and learning [13–15].

Long before contemporary metacognition research took up such questions, social psychology had made this problem central. How people come to know what they want was at stake in classic studies of attitudes and preferences: whenever researchers asked participants to evaluate options after acting on them, they were eliciting inferences about the participants' own motivational states [16, 17]. The central discovery of this tradition was that such judgments are shaped by behavior itself. Research on cognitive dissonance has repeatedly shown that people alter their reported attitudes and preferences to become more consistent with their prior actions [18–20], an effect observed even in young children and capuchin monkeys [21]. From early on, this phenomenon invited an inferential interpretation. Bem's self-perception theory [22] proposed that people infer their own attitudes from their behavior much as an external observer would infer the attitudes of another person, and recent accounts have begun to make this reading formal, describing post-choice rationalization as belief updating rather than simply motivated distortion [23–25]. Yet the computation by which this occurs has remained less clear: behavior is said to

produce dissonance, reveal preferences, or motivate rationalization, without a precise specification of how different sources of evidence are combined to determine what an individual ultimately believes they wanted. One reason is the absence of suitable measurement: classic paradigms elicit a single report after a single act, offering no way to control the evidence available for inferring one's own goals, nor to quantify and model how that evidence is combined.

The second-order framework of metacognition connects this problem to a literature in which such specifications are already routine. Whereas traditional accounts treated metacognitive judgments as direct readouts of the first-order processes that generated behavior, second-order theories distinguish an "actor" that produces decisions from a "rater" that must infer the actor's internal states from imperfect evidence — an inference cast in explicitly Bayesian terms, with prior beliefs about the actor's states updated by noisy evidence [6, 26]. On this view, self-knowledge is not privileged access but an inference problem: the observer holds a model of their own cognitive system and combines uncertain evidence to infer what state most likely generated the observed behavior — a perspective closely related to accounts of self-knowledge as self-directed mindreading [22, 25, 27–31]. Crucially, according to the second-order framework, one of the sources available to such an inference is the action itself. Confidence judgment changes after a response is executed and is influenced by motor and post-decisional signals [32–38], while judgments of agency integrate internal action signals with observable outcomes according to their reliability [39–42], analogous to reliability-weighted cue combination in perception [43]. Applied to goal inference, this framework makes the self-perception idea precise: after acting, an individual can treat their own behavior as evidence about the latent goal that produced it, integrating that evidence with prior expectations about what they were likely to want. It also

yields a sharp prediction: when external evidence about an action's apparent outcome outweighs weak internal evidence, the rational conclusion is the dissonance-like one — "apparently that is what I did; I must have wanted it."

Here, we studied this inference in a two-stage reinforcement-learning task in which, on every trial, participants acted to steer toward one of two reward-bearing options and then reported which option they had wanted. The design independently controlled each source of evidence bearing on that report: covert perturbations dissociated the performed action from its apparent outcome, the outcome display was available on only a subset of the trials, and reward volatility varied the reliability of the value-based prior [44]. Across two experiments, goal reports shifted toward the visible outcome of the action, most strongly when the prior was unreliable — exactly as reliability-weighted inference predicts. That is, the goal reports behaved as the second-order account requires: internal evidence (the value-based prior) and external evidence (the displayed outcome of the action) were combined into a single judgment, with the external source weighted more heavily exactly when the internal source was less reliable. A hierarchical Bayesian observer model reproduced the full pattern of reports with no added machinery, providing a formal, trial-level account of how people infer what they wanted from uncertain evidence about what they did.

2. Method

2.1. Participants

Participants were recruited online via Prolific and completed the task in a web browser (jsPsych; hosted on Pavlovia). All procedures were approved by the Institutional Review Board of the University of California, Los Angeles, and all participants gave informed consent. Eligibility criteria for both experiments were: age 18–40, normal or corrected-to-normal vision, English as first language, current residence in the UK or the USA, a prior Prolific approval rate of 95–100%, and at least 100 previous submissions. Participants in Experiment 1 (E1) were paid at an hourly rate of $12; participants in Experiment 2 (E2) were paid at $18 per hour plus a performance bonus of up to $6. Both experiments took approximately 50 minutes to complete.

Fifty-one participants completed E1 and 45 completed E2. We then applied a set of quality-control criteria, applied identically to both experiments, designed to remove participants who were not engaging with the task or whose data could not support the analyses (the criteria are detailed at the end of Section 2.2). Thirty participants (mean age = 35.1 yrs, 12 females) were retained in E1 and 26 (mean age = 30.8 yrs, 14 females) in E2.

2.2. Task design

In the main task, participants repeatedly steered toward one of two pairs of reward-bearing images by spinning an arrow on a wheel, reported which pair they had wanted to visit, and then chose an image from the pair they reached in order to earn coins. The task was identical in the two experiments E1 and E2, which differed only in how the goal report was collected (see Goal report below). Three features of the task — a covert perturbation of the arrow, intermittent

display of its landing position, and volatility of the reward contingencies — controlled the evidence available for this goal report, and are described in detail below. Formally, each of the 192 main trials proceeded through three stages: an action stage performed on a spin wheel, a report of the intended goal, and a reward stage consisting of a choice between two images followed by reward feedback (Figure 4-1).

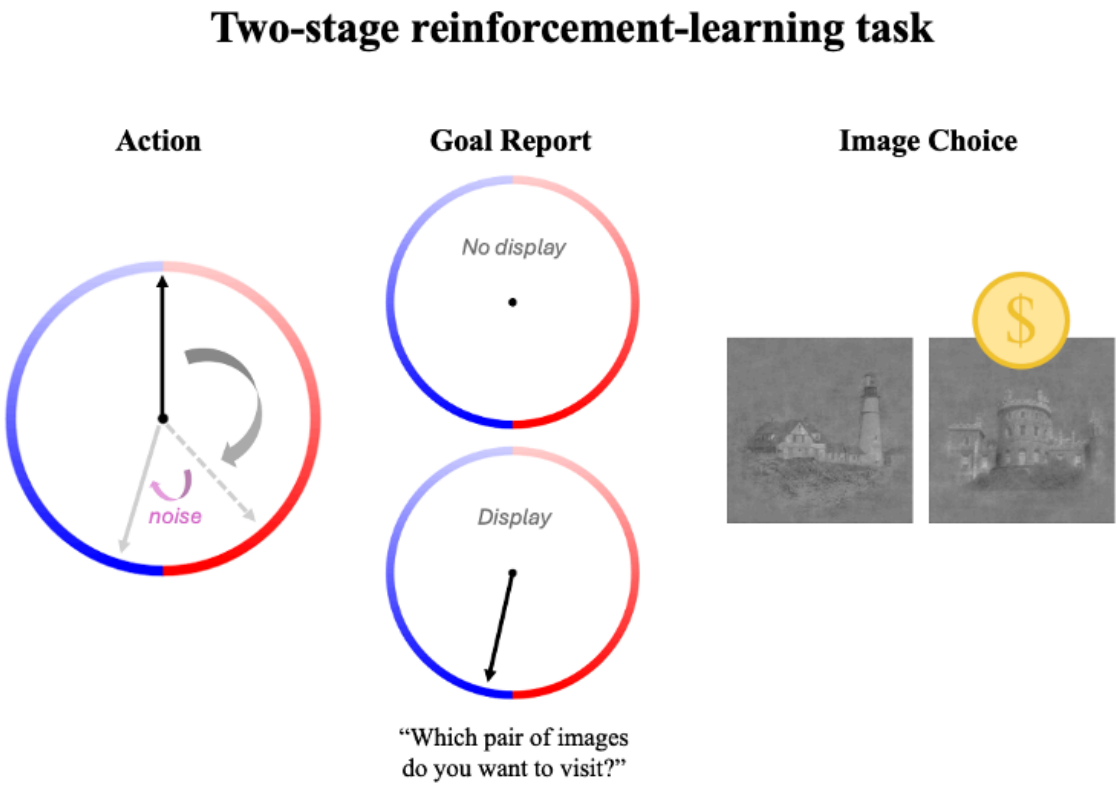

Figure 4-1. Two-stage reinforcement-learning task. **First panel**: in the action stage, participants spun an arrow clockwise around a two-colored wheel by holding and releasing the spacebar; covert noise then perturbed the release angle, and the landing position determined the probability of visiting each image category. **Second panel**: after an outcome epoch in which the final arrow position was displayed on two-thirds of trials (top: no-display trial; bottom: display trial), participants reported which pair of images they wanted to visit. **Third panel**: participants then chose between two images of the visited category (the scenes pair is shown as an example). The chosen image yielded a coin or nothing. Across blocks, the better-paying category remained fixed (stable blocks) or reversed repeatedly (volatile blocks).

Action stage (Figure 4-1, first panel). Each trial began with a central fixation cross (1,000 ms), after which the wheel appeared. The wheel was a circle divided into two colored halves, red and blue, whose left/right arrangement was randomized on every trial. Each color was associated

with one of the two image categories (bodies or scenes), and the color-to-category mapping was counterbalanced across participants. At the start of each trial, an arrow rested at a fixed position at the top of the wheel, on the upper boundary between the two halves. Participants pressed and held the spacebar to set the arrow spinning clockwise from this position, and released the spacebar to stop it; the arrow advanced in proportion to the duration of the hold, at a rate of one half-revolution per 500 ms, up to a maximum of one full revolution. Critically, the arrow was not visible during the hold period, so participants had to time their release on the basis of internal estimates alone, without online visual feedback. Precise control over the landing position was therefore challenging: a full revolution took only one second, so small errors in release timing translated into large angular displacements. The angle of the arrow at the moment of release constituted the action.

Transition structure. The mapping from the arrow's landing position to the visited image category was probabilistic. The probability that the category of the landing side would be visited increased linearly from .5 at the top of the wheel to 1.0 at the boundary at the bottom of the wheel, diametrically opposite the starting position (hereafter, the boundary). This transition rule was conveyed to participants through an explicit instruction phase; in addition, the wheel carried a color gradation that deepened toward the boundary, added to make the transition probability visually intuitive. Releasing the arrow close to the boundary therefore afforded the greatest control over which pair of images would be visited, at the cost of requiring precise timing.

Covert displacement. Unbeknownst to participants, the registered landing position did not coincide with the release position. On every trial, noise drawn from a skew-normal distribution (scale 0.5 rad, shape parameter 0.5, with the skew oriented toward the boundary) was added covertly to the release angle. A skewed rather than symmetric distribution was used so that the

noise would tend to carry the arrow toward, and at times past, the boundary rather than dispersing it evenly in both directions. As a result, on approximately one third of trials the noise displaced the arrow across the boundary between the two halves, so that it landed on the side opposite the release — dissociating the action that was performed from the outcome that the action appeared to produce. This rate presupposes that releases are aimed reasonably close to the boundary: the noise parameters were selected through simulation under this assumption, and the realized displacement rate in the actual data confirmed it (Section 3.1). Participants were never informed that such displacements existed, and all instructions treated the landing position as the direct result of the participant's own spin.

Outcome display. The final landing position was displayed on two-thirds of the trials (hereafter, *display trials*; 32 of every 48-trial block, randomly interleaved within blocks): after the release, the arrow appeared at its final position and remained visible for 1,500 ms (Figure 4-1, second panel, bottom). On the remaining third of trials (*no-display trials*), the wheel reappeared without the arrow for the same duration (Figure 4-1, second panel, top). External evidence about the apparent outcome of the action was therefore available only on display trials, whereas on no-display trials participants could rely only on internal evidence concerning their own release.

Goal report (Figure 4-1, second panel). After the action stage and the ensuing outcome epoch, participants were asked, "Which pair of images do you want to visit?" In E1, they responded by clicking one of two buttons labeled with the category names (i.e., scenes or bodies), whose positions were randomized on every trial. In E2, they responded on a four-point scale composed of the two categories crossed with two strength labels ("Definitely" and "Moderately"), presented as four buttons in a single row with the left/right assignment of categories randomized on every trial; this format yielded a graded rather than binary goal report. Responses were self-paced. On

one-fifth of trials — selected at random and not identifiable to the participant — it was guaranteed that the reported category would be visited regardless of where the arrow had landed. Participants were informed that such "bonus trials" existed but could not identify them. This arrangement made the report consequential on every trial: because any given report could determine which pair was visited, answering it engaged a genuine judgment about the currently desired goal rather than an inconsequential response.

Image choice and reward earning (Figure 4-1, third panel). Two images belonging to the visited category then appeared side by side, in randomized positions, and participants clicked one of them, revealing either a coin or an empty outcome (1,200 ms). The two images within a category carried identical reward probabilities, so that earnings depended on reaching the currently better category rather than on the choice within a pair: images of the better category were rewarded with probability .75, and images of the other category with probability .25. In E2, each coin contributed $0.05 toward the performance bonus, up to a maximum of $6; in E1, payment was not performance-contingent.

Reward volatility. The 192 trials were organized into four blocks of 48. In stable blocks, the better-paying category remained constant throughout the block; in volatile blocks, the better category reversed every 12 trials, producing three reversals per block. This manipulation controlled how certain participants could be about their own current goal: under stable contingencies, the currently better category — and hence the goal one is likely to hold — is well defined, whereas under volatility it is itself uncertain. Stable and volatile blocks alternated, and whether the session began with a stable or a volatile block was counterbalanced across participants; crossed with the color-to-category mapping, this yielded four between-participant

conditions. Block boundaries and block types were never signaled, so any sensitivity to volatility had to arise from the participants' own experience of the reward contingencies.

Training and data-quality criteria. Because the experiment was conducted online without experimenter supervision, we took particular care that participants understood the task. Before the main trials, they completed a careful, step-by-step instruction sequence with embedded practice for every element of the task, followed by multiple-choice comprehension checks with corrective feedback; further checks were administered after the main trials. Participants were informed that brief pauses were permitted during the main trials, provided that the accumulated pause time did not exceed five minutes.

To the submitted data, we applied strict data-quality criteria to ensure that only attentive, comprehending participants entered the analyses. The final samples were determined by five criteria, specified identically for the two experiments and applied to all completed sessions: (1) failure on any post-task comprehension check (10 in E1; 7 in E2); (2) release of the arrow on the same physical side on more than 80% of trials (16 in E1; 14 in E2); (3) goal reports on the same button position on more than 80% of trials (1 in E1; 2 in E2); (4) a response-time floor on the goal report, defined as a median below 400 ms or more than 10% of reports below 300 ms (0 in E1; 0 in E2);; and (5) more than five minutes of accumulated pause time during the main trials (6 in E1; 5 in E2). Because some participants failed more than one criterion, the per-criterion counts sum to more than the total numbers excluded (21 in E1; 19 in E2).

2.3. A Bayesian self-inference model

We modeled the goal report as Bayesian inference over one's own goal, performed by an observer whose generative model contains no covert displacements: because participants were never informed that displacements existed, the model treats the displayed arrow as an honest, nearly noiseless record of the observer's own action. On each trial, the observer combines three sources of evidence about the goal — a prior derived from learned values, an internal (proprioceptive-like) trace of the release, and, on display trials, the externally displayed landing position — and reports the goal with the highest posterior probability, up to response noise. The model has nine free parameters, each attached to one component of this computation; we describe the components in turn.

Value learning. The observer tracks the value of each of the four images. Let V(c,i,t) denote the value of image i of category c on trial t, initialized at 0.5. After choosing image i of category c and observing reward r(t) ∈ {0, 1}, the chosen image's value is updated by a prediction-error rule with separate learning rates for rewarded and unrewarded outcomes,

$$V_{c,i,t+1} = V_{c,i,t} + \alpha_r(r_t - V_{c,i,t}) \qquad (1)$$

where α(r) equals α^+ ("learning rate, win" in Table 4-1) when r(t) = 1 and α^- ("learning rate, loss") when r(t) = 0. In addition, after unrewarded outcomes only, the values of the three unchosen images decay toward their neutral point,

$$V_{c',i',t+1} = V_{c',i',t} + \delta(0.5 - V_{c',i',t}) \qquad (2)$$

with decay rate δ ("loss-gated decay"). This loss-gated decay implements a win-stay/lose-shift-like asymmetry — a well-established strategy in reinforcement learning that is well suited to a task in which reward contingencies alternate between stable and volatile regimes.

Value-based prior. The prior over the goal is derived from the difference between the mean values of the two categories,

$$\Delta V_t = \tfrac{1}{2}(V_{1,1,t} + V_{1,2,t}) - \tfrac{1}{2}(V_{0,1,t} + V_{0,2,t}) \qquad (3)$$

which is converted into prior log-odds by a saturating readout,

$$z_t = \beta \Delta V_t + \gamma \Delta V_t^3 \qquad (4)$$

where β ("prior strength") scales the linear contribution and the negative cubic coefficient γ ("prior saturation") caps the prior at large value differences, capturing the probability-matching-like plateau observed in participants' alignment with the better category.

Internal action evidence. The observer receives a binary internal signal f(t) indicating the side of its own release, with a sign-error probability that grows as the release approaches the boundary. Let d(t) denote the angular distance between the release position and the boundary; then

$$P(f_t = m_t) = 1 - \varepsilon_p(d_t) \qquad (7)$$

with

$$\varepsilon_p(d_t) = \varepsilon_0 + (0.5 - \varepsilon_0) \exp(-k d_t) \qquad (5)$$

where ε₀ ("proprioceptive floor") is the residual error for releases far from the boundary and k ("proprioceptive gradient") governs how quickly reliability collapses as the release approaches it. At the boundary itself (d = 0), the internal signal is uninformative (error .5): releases close to the

category boundary feel ambiguous, exactly where the covert displacement is most likely to change the landing side.

External action evidence. On display trials, the observer also receives the displayed landing side a(t). Because the observer's generative model contains no displacements, the display is treated as a nearly veridical record of its own action,

$$P(a_t = m_t) = 1 - \varepsilon_a \qquad (8)$$

with a small trust parameter ε(a) ("display trust"). It is this assumption — rational for an observer who has never been told otherwise — that makes the displayed outcome of a covertly displaced action such potent evidence about one's own goal.

Evidence combination and report. The observer allows for occasional slips between goal and executed action: the intended side m(t) matches the goal with probability 1 − s ("motor slip"),

$$P(m_t = g_t) = 1 - s \qquad (6)$$

Combining the three sources and marginalizing over the unobserved action yields the posterior log-odds of the goal,

$$\Lambda_t = z_t + \log \Sigma_m P(m \mid g = 1)P(f_t \mid m)P(a_t \mid m) - \log \Sigma_m P(m \mid g = 0)P(f_t \mid m)P(a_t \mid m) \qquad (9)$$

where the arrow term P(a | m) is included only on display trials. In E1, the report is a draw from the posterior,

$$P(y_t = 1) = \sigma(\Lambda_t) \qquad (10)$$

with σ denoting the logistic function. For E2's four-point reports, the response stage is replaced by an ordered-logistic readout of the same posterior log-odds, with a per-subject strength cutpoint c and response bias b defining three category boundaries,

$$P(y_t \le j) = \sigma(\kappa_j - \Lambda_t), \quad \kappa = (b - c, b, b + c) \qquad (11)$$

so that "Definitely" responses arise when the posterior log-odds exceed the outer cutpoints. The primary E2 analyses fit the model to reports collapsed to sides, exactly as in E1, for like-for-like comparison of parameters across experiments; the ordered-logistic readout was then used to test whether the same posterior log-odds predict trial-level report strength.

All nine parameters (eleven for the ordered-logistic readout) were estimated hierarchically. Each subject-level parameter was obtained by transforming a latent Gaussian variable with population mean and standard deviation, using a non-centered parameterization, with range-appropriate transforms (e.g., inverse-logistic mappings onto bounded supports). Population means received weakly informative Gaussian priors and standard deviations half-Gaussian priors. Posteriors were sampled with the No-U-Turn Sampler in Stan [45; four chains, 600 warmup and 500 sampling iterations, adapt_delta = .95]; all fits converged with no divergent transitions. Goal-report trials entered the likelihood after response-time trimming (100–5,000 ms, then a per-subject 4-MAD criterion; 90.7% and 90.6% of trials retained in E1 and E2). Model adequacy was assessed by approximate leave-one-out cross-validation [46] and by posterior-predictive checks, in which reports simulated from the fitted model were compared with the observed data in every design cell.

3. Results

3.1. Task engagement

The two experiments differed only minimally in design — E2 replaced E1's binary goal report with a graded, four-point report — and, as will become apparent below, the pattern of behavior in E2 did not differ qualitatively from that in E1. We therefore report pooled results throughout, giving the per-experiment values in parentheses and noting specific differences explicitly only where they arise. That two samples of different participants, collected at different times, converged on the same pattern is itself a first indication that the experiment ran reliably. Before turning to our primary results, we verified that the task functioned as designed and that participants engaged with it.

Participants used the action stage as the transition rule prescribes: arrow releases were concentrated toward the boundary, where control over the visited category is greatest, with 73.3% falling within 1 rad (Figure 4-2A; mean of participants' median distances: 0.66 rad). Both sides of the wheel were used (mean majority-side proportion: 59.9%), and the realized rate of covert displacement, 31.5% of trials, matched the intended rate of approximately one-third. Release positioning was, moreover, calibrated by the displayed outcomes: after display trials on which the landing confirmed the release, the next release moved 0.066 rad closer to the boundary — the riskier, higher-control region ($t(54) = 4.73$, $p < .001$; E1: 0.087; E2: 0.041) — whereas this emboldening was absent after displaced displays (0.008 rad; interaction $t(54) = 3.82$, $p < .001$). In other words, participants calibrated their motor behavior appropriately on the basis of the displayed outcomes. Together, these patterns indicate that the action stage elicited genuinely goal-directed behavior: participants entered each trial with a category they pursued, aiming and adjusting their releases accordingly.

Participants also learned the reward contingencies (Figure 4-2B). In stable blocks, releases were directed toward the side of the currently better-paying category well above chance ($t(55) = 6.95$, $p < .001$; E1: 59.8%; E2: 60.7%). In volatile blocks this alignment was reduced ($t(55) = 4.14$, $p < .001$; E1: 52.9%; E2: 56.2%), and the stable-volatile difference (+5.7 percentage points, $t(55) = 3.59$, $p < .001$; E1: +6.8; E2: +4.5) confirms that the volatility manipulation degraded value-based steering as intended. Together, these checks establish that participants understood the task, pursued reward, and tracked the contingencies — the preconditions for interpreting the goal reports, our primary measure of interest, to which we now turn.

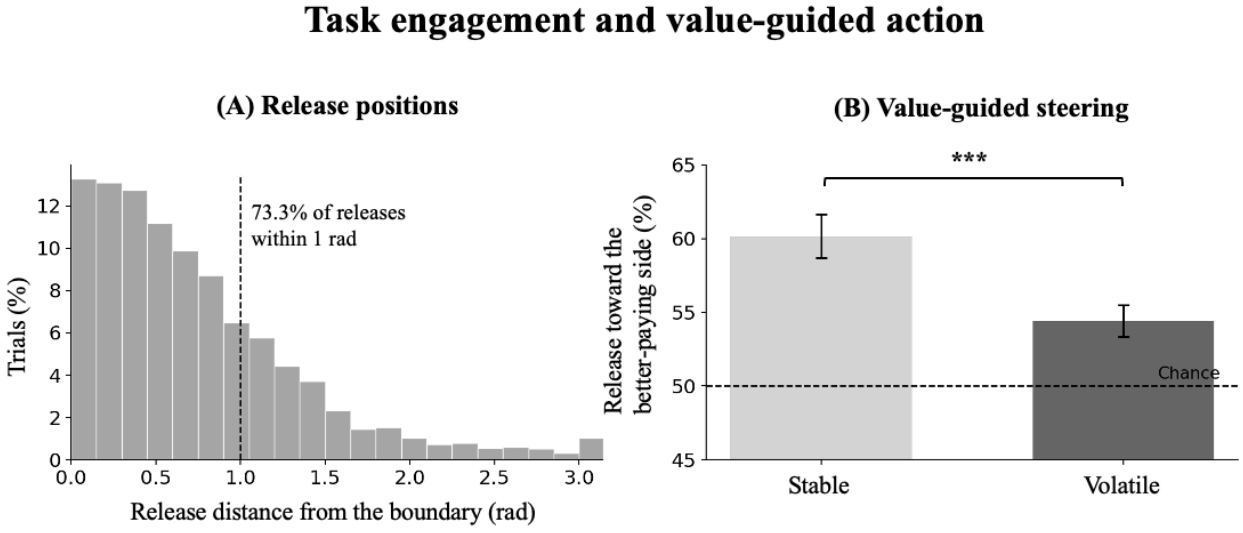

Figure 4-2. Task engagement and value-guided action, pooled across experiments (N = 56). **(A)** Distribution of release distances from the boundary over all trials of both experiments. Releases concentrated toward the boundary, where control over the visited category is greatest; 73.3% fell within 1 rad (dashed line). **(B)** Proportion of releases directed toward the side of the currently better-paying category (mean of participant means; error bars, SEM) in stable and volatile blocks. Steering exceeded chance (dashed line) in both block

types and was reduced under volatility, confirming that participants pursued reward and that the volatility manipulation degraded value-based steering as intended.

3.2. Goal reports adopt the visible outcome of one's own action

Participants' goal reports tracked the displayed arrow far beyond what their actions warranted (Figure 4-3A). Across both experiments, goal reports matched the arrow's final landing side on 78.9% of display trials versus 64.9% of no-display trials — a display effect, i.e., an increased alignment of the reported goal with the apparent outcome when the action's outcome is observed ($t(55) = 7.96$, $p < .001$; E1: +17.5, $t(29) = 7.37$, $p < .001$; E2: +9.9, $t(25) = 4.12$, $p < .001$). The critical test comes from displaced trials: trials on which the covert noise carried the arrow across the boundary, so that the final landing position lay on the side opposite to the one toward which the participant had actually released the arrow (Figure 4-3B). On these trials, the apparent outcome of the action conflicted with the action actually performed. When the landing position was not displayed, reports went with the action: only 34.9% of reports matched the displaced landing side — that is, most participants reported the side of their actual release. When the landing position was displayed, this proportion nearly doubled, to 64.8% (+29.9 points, $t(54) = 9.60$, $p < .001$): seeing the displaced arrow led participants to report the side of the apparent outcome rather than the side of their actual release. In other words, the visible outcome served as external evidence strong enough to override the internal estimate of one's own release: when the world's version of the action conflicted with the body's, the majority of reports followed the world's — the dissonance-reduction signature, produced trial by trial in a value-based task.

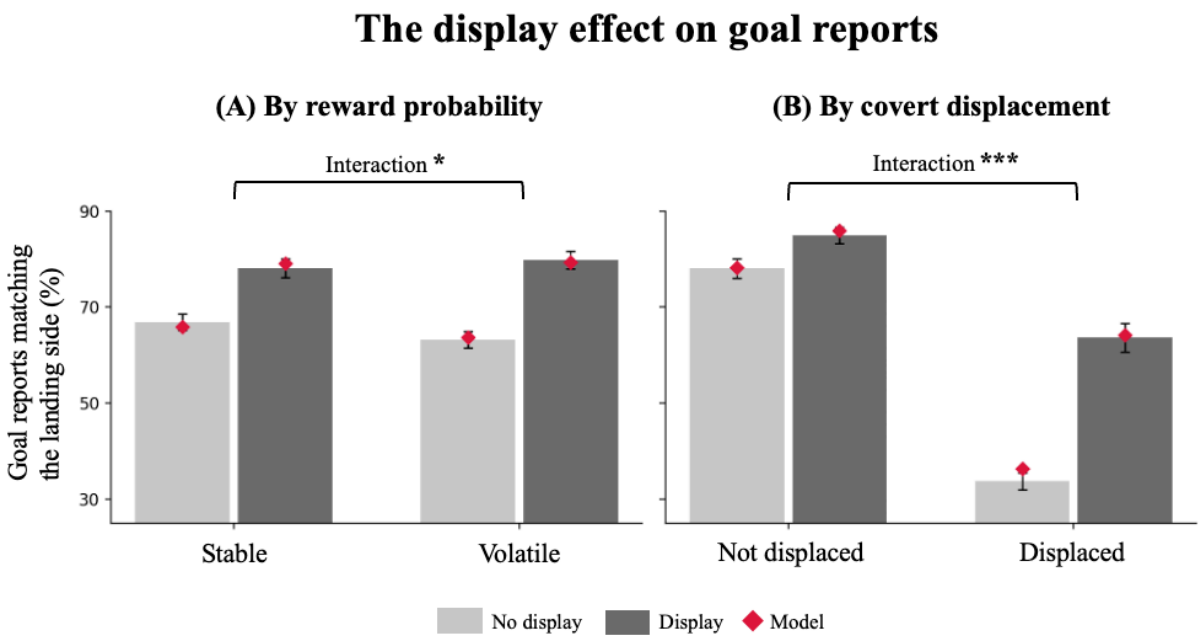

Figure 4-3. Goal reports adopt the displayed outcome of the action, pooled across experiments (N = 56). Bars show the proportion of goal reports matching the arrow's final landing side (mean of participant means; error bars, SEM) on no-display (light) and display (dark) trials; red diamonds show the posterior-predictive means of the computational model fitted separately to each experiment. **(A)** Split by reward volatility. The display effect — the increase in outcome-matching when the landing position was displayed — was larger in volatile blocks, in which the value-based prior over one's own goal was less reliable. **(B)** Split by covert displacement. On displaced trials, the noise carried the arrow to the side opposite the participant's actual release; displaying the landing position roughly doubled the rate at which reports followed the apparent outcome rather than the actual release. The pattern was qualitatively identical in the two experiments; per-experiment statistics are reported in the text.

3.3. Adoption scales with the uncertainty of internal evidence

If the goal report is an inference that combines multiple informative cues in proportion to their reliability, external evidence should matter most when internal evidence is weak. Volatile blocks degrade the value-based prior — the participant's own best internal grounds for knowing what they wanted — and, as predicted, the display effect was reliably larger in volatile than in stable blocks (Figure 4-3A): interaction +5.3 points, $t(55) = 2.64$, $p = .011$ (E1: +5.9, $t(29) = 2.22$, $p = .034$; E2: +4.6, $t(25) = 1.48$, $p = .15$). The interaction was, moreover, composed exactly as the account predicts: under volatility, matching fell on no-display trials, where the degraded prior is

the only guide (66.8% to 63.2%), and rose on display trials, where the external evidence gains weight (78.1% to 79.8%). This interaction is the signature prediction of the second-order account of goal inference: if judgments about one's own goals are computed by combining internal and external evidence in proportion to their reliability, then degrading the internal evidence must increase the weight given to the external evidence. The volatility manipulation did precisely this, and the goal reports followed — demonstrating formally, at the level of trial-by-trial behavior, that goal reports arise from evidence combination rather than from a privileged readout of one's own intentions.

3.4. A Bayesian self-inference model explains the choice pattern

The nine-parameter Bayesian self-inference model reproduced the full choice pattern in both experiments (Figure 4-3, red diamonds): all eight design cells (display x volatility and display x push) fell within the posterior-predictive intervals in E1 and in E2 (all cell-level posterior-predictive $p >= .11$), including the near-chance following of hidden pushes and the strong adoption of displayed ones. Out-of-sample accuracy was substantial in both experiments (per-trial elpd -0.38 and -0.35 against a chance value of -0.69). The model also reproduced the direction of the uncertainty modulation — a larger display effect under volatility — though at reduced magnitude, indicating that the rational core captures most, but not all, of the interaction.

Parameter estimates (Table 4-1) tell a mechanistically specific story. The perceptual and motor machinery was numerically indistinguishable across experiments — most notably the proprioceptive reliability gradient (k = 2.93 vs. 2.94) — while display trust differed by a factor of three (arrow noise 0.034 vs. 0.098), localizing E2's attenuated display effects to a single

interpretable parameter: eliciting graded reports made participants weight the external display less, leaving the rest of the self-inference intact. We return to this point in the Discussion.

Parameter	Interpretation	E1	E2
learning rate (win), α^+	value update after reward	0.195	0.218
learning rate (loss), α^-	value update after no reward	0.162	0.132
prior strength, β	value difference to goal prior	6.85	8.48
prior saturation, γ	nonlinearity of the prior readout	-6.62	-1.89
motor slip, s	goal-action slip rate	0.078	0.071
proprioceptive floor, ε_0	felt-release error, far from midpoint	0.058	0.050
proprioceptive gradient, k	reliability decline toward midpoint	2.93	2.94
display trust (arrow noise), ε_a	distrust of the displayed arrow	0.034	0.098
loss-gated decay, δ	forgetting of unchosen values after loss	0.053	0.155

Table 4-1. Posterior mean population parameters of the self-inference model in the two experiments. Symbols follow Equations 1–8. Perceptual-motor parameters transfer across report formats; display trust does not.

3.5. Graded reports corroborate the inference account (E2)

E2's graded goal reports, which extended E1's binary format, allowed a direct test of the inferential account of goal judgment. The account is committed to a specific internal quantity: on every trial, the observer arrives at a posterior probability that a given category was its goal, and the report is a readout of this posterior. If participants indeed employ this mechanism, the magnitude of the inferred posterior should be visible in their graded reports — trials on which the posterior is extreme should attract "definitely" responses, and trials on which it is intermediate should attract "moderately." We therefore refit the model to the four-point reports, replacing only the response stage with an ordered-logistic readout (Equation 11) and adding no strength-specific mechanisms: whether "definitely" or "moderately" is chosen is determined entirely by where the same posterior log-odds that drive the choice fall relative to the participant's cutpoints. This stripped-down readout reproduced the graded-report data (Figure 4-4). The behavioral effects were small and did not reach significance in subject-level tests on their

own: participants answered "definitely" somewhat more often when the landing position was displayed (P("definitely") = .67 vs. .64; $t(25) = 1.42$, $p = .17$) and somewhat less often in volatile blocks (.63 vs. .69; $t(25) = 1.77$, $p = .09$). The model-based test, however, does not rest on these aggregate contrasts: fitted to the trial-level reports, the ordered-logistic readout predicted both effects quantitatively, including their small size (display: observed +0.034, predicted +0.035, posterior-predictive $p = .96$; volatility: observed -0.064, predicted -0.053, $p = .33$). The modest magnitude of the observed effects is thus not a shortfall of the data relative to the account — it is what the account predicts — and both effects fall where the magnitude of the same inferred posterior that produces the choices places them.

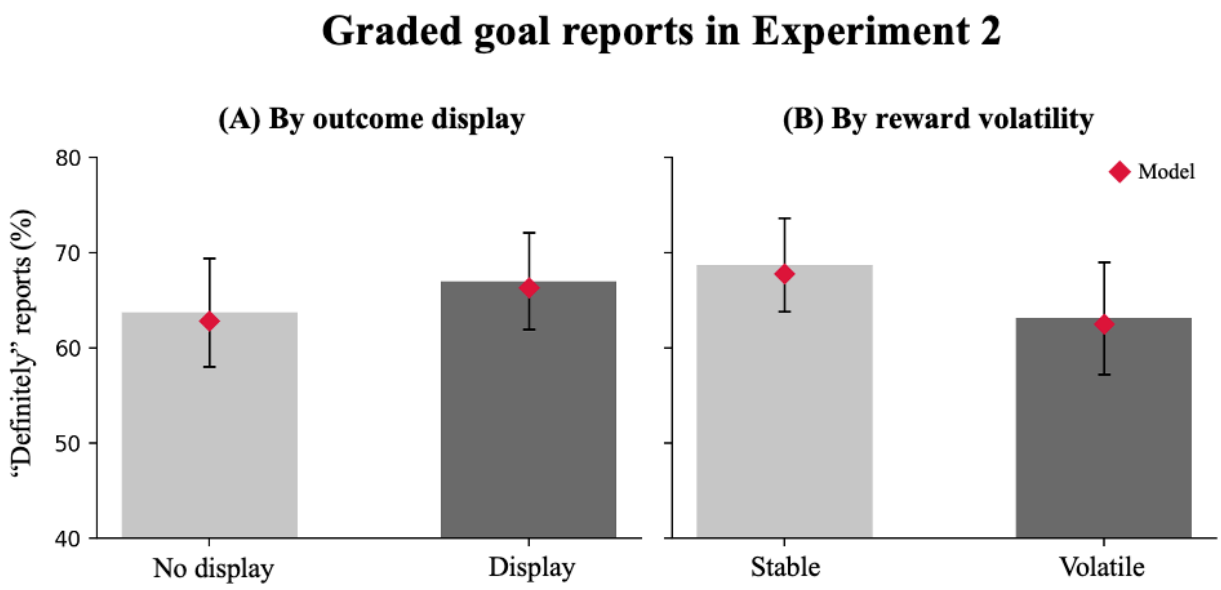

Figure 4-4. Report strength tracks the inferred posterior (Experiment 2, n = 26). Bars show the proportion of "definitely" (vs. "moderately") reports (mean of participant means; error bars, SEM); red diamonds show posterior-predictive means of the model fitted to the four-point reports with an ordered-logistic readout and no strength-specific mechanisms. **(A)** By outcome display: the proportion of "definitely" reports is higher when the landing position is displayed. **(B)** By reward volatility: it is lower in volatile blocks. Both effects are quantitatively reproduced by the magnitude of the same posterior that generates the choices.

3.6. Post-conflict slowing of instrumental choice

The action-based model of cognitive dissonance [47] offers a complementary, process-level perspective on these findings. On this account, the function of dissonance processes is to support effective action: conflict between action-related cognitions — here, between what one did and what one wanted — produces an aversive state that interferes with subsequent goal-directed behavior until the conflict is resolved, for example by bringing the goal into line with the action. Applied to the present task, the account makes a further prediction beyond the goal reports themselves: on trials in which the conflict between the displayed outcome and the performed action remains unresolved by bringing the goal report to the displayed outcome, the next goal-directed act — the image choice — should be slowed.

Consistent with such a conflict-specific cost, in E1 the second-stage image choice was slowed after reports that conflicted with a displayed, displaced arrow (Figure 4-5). On displaced trials, choices were 135 ms slower following conflicting reports when the contradiction was visible on screen ($t(27) = 3.34$, $p = .003$), compared with only 59 ms slower when it was not displayed; likewise, participants who resisted the displayed arrow — reporting the side of their actual release — were 113 ms slower than those who adopted it ($t(21) = 2.01$, $p = .057$). The pattern thus pointed consistently in the predicted direction: the slowing was larger whenever the conflict was actually visible to the participant. The decisive comparison, however — the difference between the slowing on display trials and on no-display trials, which isolates the component that depends on perceiving the conflict — did not reach significance (+66 ms, $p = .15$), and no comparable pattern appeared in E2, possibly because the four-option report format softened the binary commit-or-resist structure that the action-based account requires. We therefore treat the

post-conflict slowing as a preliminary trend motivating a purpose-built test, rather than as an established result.

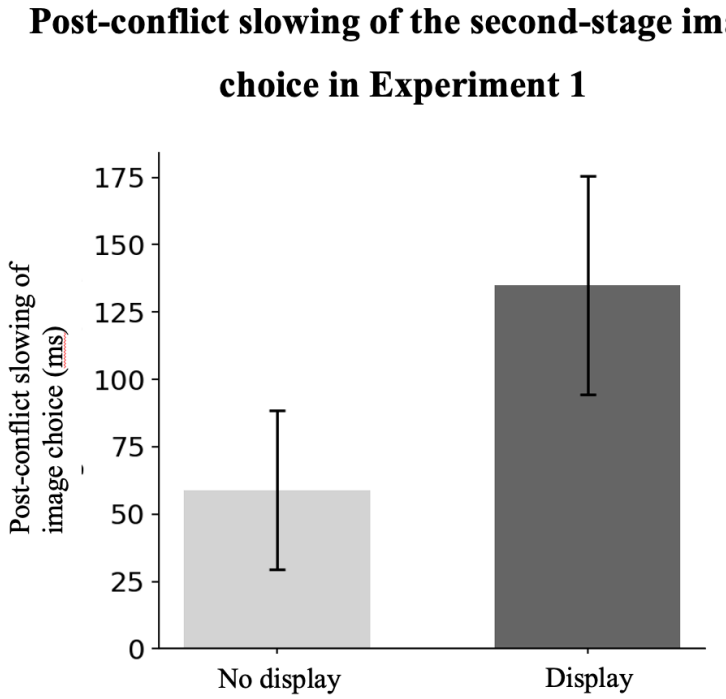

Figure 4-5. Post-conflict slowing of the second-stage image choice (Experiment 1, n = 30). Bars show the increase in image-choice response time following goal reports that conflicted with the apparent outcome, relative to reports consistent with it, on displaced trials (mean of participant means; error bars, SEM). The slowing was reliable when the landing position was displayed (+135 ms) and smaller when it was not (+59 ms), but the difference between the two — the component specific to perceiving the conflict — did not reach significance (see text).

3.7. Does the goal inference write back into value?

If the goal report is an inference, a further question arises: once expressed, does the inference feed back into the value system?

A model-based test added a single feedback parameter κ to the self-inference model: after each report, the values of the reported category's images were nudged upward by κ. Fitted alone, κ

was strongly positive in both experiments (E1: +0.25, 95% CI [.20, .29]; E2: +0.24 [.18, .29]) and improved out-of-sample fit substantially (+104 and +84 elpd over the base model). This result, however, proved to be confounded: a control model that replaced value feedback with a value-free report habit — an exponential trace of previous reports entering the report stage directly, leaving values untouched — matched these gains in both experiments. Report perseveration can masquerade as revaluation, because in a report-only likelihood the two mechanisms occupy nearly the same niche. When both mechanisms were fitted jointly, κ remained credibly positive in E1 (+0.17 [.06, .26]), its predictive advantage concentrated precisely where a value-mediated mechanism predicts — on trials immediately following a visit to the reported category that was rewarded, the signature of reward-consolidated commitment — and a reward-blind two-timescale habit could not reproduce this pattern. In E2, by contrast, the same test was inconclusive: κ's interval included zero ([−.01, .26]), it added no predictive advantage over the habit model, and the reward-coupled signature did not reappear.

The asymmetry between experiments parallels a pattern seen throughout this paper: E2's graded report format attenuated every commitment-related effect — the display effect, display trust (Section 3.4), and the post-conflict slowing (Section 3.6). If value write-back is driven by committing to a single answer, the binary format of E1 is where it should be strongest, and E2 constitutes a test at a weaker dose. This account is plausible in light of our format findings, which independently show weakened commitment effects under graded reports; but those findings concern other measures, and the κ estimates themselves did not differ detectably between experiments, so format moderation of the write-back remains an interpretation rather than a demonstrated result.

4. Discussion

Knowing what one wanted is not as simple as it feels. Across two experiments, we placed the goal of an action under experimental control of its evidence: covert displacements dissociated what participants did from what they appeared to do, intermittent display controlled whether the apparent outcome was observable, and reward volatility varied the reliability of the value-based prior. Goal reports behaved exactly as a reliability-weighted inference should. Participants adopted the visible outcome of their own action — on displaced trials, displaying the arrow nearly doubled the rate at which reports followed the apparent outcome over the actual release — and adoption grew when the prior was degraded by volatility. A nine-parameter Bayesian observer, which treats the report as a posterior over one's own goal given the learned values, a proprioceptive trace of the release, and the displayed outcome, reproduced every design cell in both experiments; in E2, the strength of graded reports was read out from the same posterior with no additional machinery. Together, the results establish goal inference as a metacognitive judgment that can be measured, manipulated, and modeled trial by trial.

These findings extend metacognition research along the axis that recent commentaries have urged [6, 7]: from evaluating the reliability of first-order processes to inferring their content. The second-order framework was developed to explain how a "rater" infers the accuracy of an "actor's" decision from imperfect evidence, chiefly in the domain of perceptual decision making [26]; the present results carry the same insight into value-based decision making, where the latent states to be inferred — one's own goals and preferences — are exactly those that the classic literatures on attitude and preference change made central. In this regime, the same

architecture supports inference about what the actor was trying to do, and it does so quantitatively, not merely qualitatively — the interaction between display availability and prior reliability is the signature of reliability-weighted cue combination [43], and it appeared in both experiments. The extension thus connects two lines of research on self-knowledge that have developed largely independently — the psychophysical tradition of metacognition and the classic study of how people come to know what they want. The goal-inference account also joins judgments of agency, which combine internal action signals with observed outcomes according to their reliability [41, 42], and post-decisional revisions of confidence [32, 35] in a common inferential family, while targeting a latent state — the goal — that neither of those literatures measures. Importantly, the results also give concrete empirical content to accounts of self-knowledge as self-directed mindreading [22, 27, 31]: our participants treated evidence about their own action much as an observer would treat evidence about another person's, to the point that when the world's version of the action contradicted the self's, the world usually won.

For cognitive dissonance, the contribution is a process model of the adjustment itself. The signature phenomenon — a reported preference moving into line with prior action — emerged trial by trial in a value-based task, was turned on and off by controlling a single evidence source, and was captured quantitatively by a normative Bayesian observer containing no motivational construct — the adjustment itself required neither an aversive state nor a consistency drive, only Bayesian evidence combination by an observer whose generative model does not include the experimenter's covert manipulation. This gives formal expression to the classic self-perception insight [22] and joins recent treatments of rationalization as belief updating [24, 25], but with a trial-level likelihood in place of a redescription. Importantly, this inferential reading does not reject the original dissonance claim that discomfort arises from unresolved conflict; it suggests,

rather, that the inference may be what generates the discomfort. If the goal is a belief maintained by evidence, then a conflict between what one did and what one wanted is a violated inference — and violated or unresolved inference appears to be aversive in its own right: inconsistencies and expectancy violations elicit a common aversive arousal that motivates compensation [48], and dissonance reduction has been recast as the minimization of prediction error in predictive-processing terms [49]. Neurally, dissonance engages prefrontal conflict circuitry — in particular the dorsal anterior cingulate cortex, a region centrally implicated in detecting and resolving conflict that is experienced as aversive [50] — and dissonance-related activity there predicts subsequent attitude change [23, 51]. On this view, the inferential and the motivational accounts of dissonance — often pitted against each other — describe two faces of a single mechanism: the inference specifies what the conflict is and how it is resolved, while the discomfort signals that it is not yet resolved. The mechanism's simplicity speaks to the phylogenetic and developmental breadth of dissonance-like effects [21]: reliability-weighted cue combination requires no sophisticated introspective apparatus, which is precisely what one expects of a computation shared with young children and nonhuman primates. Beyond these theoretical considerations, the current task design also sidesteps the selection-artifact critique of classic free-choice paradigms [52]: here the conflict between action and apparent outcome is created by the experimenter, not selected by the participant's own noisy preferences, so the observed adjustment cannot be an artifact of conditioning on the choice.

The one parameter that failed to transfer between experiments carries its own lesson. When the report format changed from binary to graded, the entire perceptual-motor machinery of the inference was numerically unchanged — most strikingly the proprioceptive reliability gradient (2.93 vs. 2.94) — while trust in the displayed outcome fell by a factor of three. How one is asked

evidently changes how the evidence is weighed: a graded scale offers a way to register partial conviction, and participants took it, discounting the external cue rather than committing to it. This is measurement reactivity of a specific and interpretable kind. Methodologically, it warns that self-report formats are not neutral windows onto self-inference; studies eliciting reports in different formats may disagree about the potency of external evidence while measuring the same underlying computation, and a model with interpretable parameters can localize such differences to a single locus, as it did here. Theoretically, the direction of the shift is suggestive: a binary report forces full commitment to one answer, and committing to the externally supported answer may itself be part of what gives external evidence its grip. Suggestive precedents exist: eliciting metacognitive judgments is known to feed back on the first-order process it measures [53, 54], and inducing deliberate introspection can shift preference-based decisions away from immediate commitments [55]. The present results locate such reactivity in a single interpretable parameter of the inference — a point of contact with the action-based perspective considered next.

The inferential account describes how the goal moves; the action-based model of cognitive dissonance [47] proposes why it should move at all: conflict between what one did and what one wanted is costly for subsequent goal-directed action, and dissonance reduction functions to clear that cost. This proposal is seminal: it remains the most developed answer to why dissonance reduction should exist at all, and its clear elucidation would deepen our understanding of the whole family of dissonance-like phenomena. Yet it has rarely been put to rigorous test. Direct tests of the model have largely relied on neural and embodiment markers of approach motivation following commitment [56, 57] — evidence consistent with the account but not isolating its distinctive behavioral prediction, a cost of unresolved conflict in subsequent goal-directed action, on a trial-by-trial basis. The inferential and the action-based accounts are complementary —

reliability-weighted inference is a computational description of exactly the adjustment whose function the action-based model specifies — and the action-based model adds one prediction that pure inference does not make: a residual behavioral cost on trials where the conflict is left unresolved. We observed a trend with precisely this profile. Image choices slowed after goal reports that resisted a visibly conflicting outcome, and the slowing was larger when the conflict was actually on screen than when it was hidden, although the decisive difference between those two conditions did not reach significance and no counterpart appeared in E2, whose four-option format dilutes the binary commit-or-resist structure the account requires. It is at least suggestive that the same format change that softened commitment also lowered display trust: both effects are consistent with a single underlying variable — the pressure to commit to one version of the action — modulating both the weighting of evidence and the cost of resisting it. A purpose-built test, with a committing report format and adequate power in the conflicted cells, is required before the slowing can be treated as established; convergent neural evidence that dissonance engages conflict-related circuitry [23] makes the prediction worth that investment.

Several limitations bound these conclusions. The experiments were conducted online; the strict quality controls and the replication of every qualitative pattern across two independently collected samples mitigate, but do not remove, the case for laboratory confirmation. The uncertainty modulation, though reproduced in direction by the model, exceeded its rational magnitude by roughly a factor of two in both experiments, implying an amplification of external evidence under uncertainty that static reliability weighting does not capture. Most fundamentally, the model's internal-evidence channel rests on a latent variable: the participant's actual motor intention on each trial is never observed but inferred jointly with everything else, so the claim

that reports override a definite internal estimate — rather than fill in an absent one — is supported only indirectly, through the model comparison.

A future experiment might address this central limitation by bringing the task into the laboratory and using eye tracking to obtain a trial-level, report-independent measure of motor intention. Because the wheel is visible during the hold while the arrow is not, the region a participant monitors while timing the release could provide an indication of the location toward which the action is being steered, measured before the covert displacement and before any outcome appears. Such a measure could make the intended side, currently marginalized in the posterior (Equation 9), directly observable and thereby sharpen estimation of the internal-evidence channel and downstream model parameters. It could also allow conflict to be defined by intention rather than by release, making it possible to test whether adoption of the displayed outcome and subsequent slowing are specifically associated with trials in which the participant had clearly intended the opposite side. Finally, pre-release gaze dynamics could help clarify the amplified display effect under uncertainty: if this effect reflects competition between candidate goals within a trial, such competition might be expressed as gaze alternation between the two target regions. Thus, laboratory studies incorporating independent measures of motor intention could help resolve two questions left open by the present experiments: the source of the amplified uncertainty interaction and the status of the dissonance-related behavioral trend.

People speak as though their goals were transparently known to them. The present results indicate that, in value-guided action under uncertainty, knowing what one wanted is instead an inference — assembled from learned values, a fallible sense of one's own movement, and whatever the world displays about the action — and that this inference obeys the same reliability-weighting norms as inference about the external world. Seen through this

metacognitive lens, a question with a classic history — how people come to know what they want — becomes continuous with the question of how they know anything else: by inference from evidence, weighted by reliability, and now observable one trial at a time.

References

1. Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. Frontiers in Human Neuroscience, 8, 443.

2. Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. Neuron, 88, 78-92.

3. De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. Nature Neuroscience, 16, 105-110.

4. Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. Nature Neuroscience, 18, 1159-1167.

5. Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2017). Explicit representation of confidence informs future value-based decisions. Nature Human Behaviour, 1, 0002.

6. Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. Annual Review of Psychology, 75, 241-268.

7. Nakamura, T., & Lau, H. (2025). Perceptual metacognition beyond confidence. Neuron, 113(15), 2377-2379.

8. Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), Classical conditioning II (pp. 64-99). Appleton-Century-Crofts.

9. Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience, 9, 545-556.

10. Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231-259.

11. Johansson, P., Hall, L., Sikstrom, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. Science, 310, 116-119.

12. Hall, L., Johansson, P., Tarning, B., Sikstrom, S., & Deutgen, T. (2010). Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. Cognition, 117, 54-61.

13. Sharot, T., Velasquez, C. M., & Dolan, R. J. (2010). Do decisions shape preference? Evidence from blind choice. Psychological Science, 21, 1231-1235.

14. Cockburn, J., Collins, A. G. E., & Frank, M. J. (2014). A reinforcement learning mechanism responsible for the valuation of free choice. Neuron, 83, 551-557.

15. Chambon, V., Thero, H., Vidal, M., Vandendriessche, H., Haggard, P., & Palminteri, S. (2020). Information about action outcomes differentially affects learning from self-determined versus imposed choices. Nature Human Behaviour, 4, 1067-1079.

16. Cooper, J. (2007). Cognitive dissonance: Fifty years of a classic theory. Sage.

17. Harmon-Jones, E., & Mills, J. (2019). An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In E. Harmon-Jones (Ed.), Cognitive dissonance: Reexamining a pivotal theory in psychology (2nd ed., pp. 3-24). American Psychological Association.

18. Festinger, L. (1957). A theory of cognitive dissonance. Stanford University Press.

19. Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. Journal of Abnormal and Social Psychology, 58, 203-210.

20. Brehm, J. W. (1956). Postdecision changes in the desirability of alternatives. Journal of Abnormal and Social Psychology, 52, 384-389.

21. Egan, L. C., Santos, L. R., & Bloom, P. (2007). The origins of cognitive dissonance: Evidence from children and monkeys. Psychological Science, 18, 978-983.

22. Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. Psychological Review, 74, 183-200.

23. van Veen, V., Krug, M. K., Schooler, J. W., & Carter, C. S. (2009). Neural activity predicts attitude change in cognitive dissonance. Nature Neuroscience, 12, 1469-1474.

24. Gershman, S. J. (2019). How to never be wrong. Psychonomic Bulletin & Review, 26, 13-28.

25. Cushman, F. (2020). Rationalization is rational. Behavioral and Brain Sciences, 43, e28.

26. Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. Psychological Review, 124, 91-114.

27. Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. Behavioral and Brain Sciences, 32, 121-138.

28. Frith, C. D. (2012). The role of metacognition in human social interactions. Philosophical Transactions of the Royal Society B, 367, 2213-2223.

29. Proust, J. (2013). The philosophy of metacognition: Mental agency and self-awareness. Oxford University Press.

30. Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. Trends in Cognitive Sciences, 18, 186-193.

31. Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing ourselves together: The cultural origins of metacognition. Trends in Cognitive Sciences, 24, 349-362.

32. Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-specific disruption of perceptual confidence. Psychological Science, 26, 89-98.

33. Siedlecka, M., Paulewicz, B., & Wierzchon, M. (2016). But I was so sure! Metacognitive judgments are less accurate given prospectively than retrospectively. Frontiers in Psychology, 7, 218.

34. Gajdos, T., Fleming, S. M., Saez Garcia, M., Weindel, G., & Davranche, K. (2019). Revealing subthreshold motor contributions to perceptual confidence. Neuroscience of Consciousness, 2019(1), niz001.

35. Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. Nature, 461, 263-266.

36. Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. Psychological Review, 117, 864-901.

37. van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. eLife, 5, e12192.

38. Wokke, M. E., Achoui, D., & Cleeremans, A. (2020). Action information contributes to metacognitive decision-making. Scientific Reports, 10, 3632.

39. Metcalfe, J., & Greene, M. J. (2007). Metacognition of agency. Journal of Experimental Psychology: General, 136, 184-199.

40. Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. American Psychologist, 54, 480-492.

41. Moore, J. W., & Fletcher, P. C. (2012). Sense of agency in health and disease: A review of cue integration approaches. Consciousness and Cognition, 21, 59-68.

42. Haggard, P. (2017). Sense of agency in the human brain. Nature Reviews Neuroscience, 18, 196-207.

43. Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415, 429-433.

44. Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. Nature Neuroscience, 10, 1214-1221.

45. Carpenter, B., Gelman, A., Hoffman, M. D., et al. (2017). Stan: A probabilistic programming language. Journal of Statistical Software, 76(1).

46. Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, 27, 1413-1432.

47. Harmon-Jones, E., Harmon-Jones, C., & Levy, N. (2015). An action-based model of cognitive-dissonance processes. Current Directions in Psychological Science, 24, 184-189.

48. Proulx, T., Inzlicht, M., & Harmon-Jones, E. (2012). Understanding all inconsistency compensation as a palliative response to violated expectations. Trends in Cognitive Sciences, 16, 285-291.

49. Kaaronen, R. O. (2018). A theory of predictive dissonance: Predictive processing presents a new take on cognitive dissonance. Frontiers in Psychology, 9, 2218.

50. Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. Trends in Cognitive Sciences, 8, 539-546.

51. Izuma, K., Matsumoto, M., Murayama, K., Samejima, K., Sadato, N., & Matsumoto, K. (2010). Neural correlates of cognitive dissonance and choice-induced preference change. Proceedings of the National Academy of Sciences, 107, 22014-22019.

52. Chen, M. K., & Risen, J. L. (2010). How choice affects and reflects preferences: Revisiting the free-choice paradigm. Journal of Personality and Social Psychology, 99, 573-594.

53. Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. Psychonomic Bulletin & Review, 10, 177-183.

54. Double, K. S., & Birney, D. P. (2019). Reactivity to measures of metacognition. Frontiers in Psychology, 10, 2755.

55. Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. Journal of Personality and Social Psychology, 60, 181-192.

56. Harmon-Jones, E., Gable, P. A., & Price, T. F. (2011). Leaning embodies desire: Evidence that leaning forward increases relative left frontal cortical activation to appetitive stimuli. Biological Psychology, 87, 311-313.

57. Harmon-Jones, E., Harmon-Jones, C., Serra, R., & Gable, P. A. (2011). The effect of commitment on relative left frontal cortical activity: Tests of the action-based model of dissonance. Personality and Social Psychology Bulletin, 37, 395-408.

58. 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, 1541-1543.

59. Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. Psychonomic Bulletin & Review, 10, 843-876.

60. Izuma, K., & Murayama, K. (2013). Choice-induced preference change in the free-choice paradigm: A critical methodological review. Frontiers in Psychology, 4, 41.

61. Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. Science, 324, 759-764.

62. Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Consciousness and Cognition, 21, 422-430.

63. Vickers, D. (1979). Decision processes in visual perception. Academic Press.

Chapter 5. General Discussion

This dissertation began from a simple requirement of adaptive cognition: internal representations must persist enough to guide behavior yet remain flexible enough to be revised, and the balance between the two is struck not in a single pass but through an iterative exchange between what the mind currently holds and what the environment provides. The three studies examined this exchange in three settings, at different timescales and levels of cognition. In perception, intracranial recordings showed that a briefly presented input is not read against a recently formed memory at a stroke: information about the correspondence between the internal representation and the current input emerged early, persisted over the better part of a second, and repeatedly informed the developing perceptual decision (Chapter 2). In decision evaluation, revision of confidence in response to feedback went beyond any single-pass weighting: feedback that conflicted with a confident judgment prompted a renewed consultation of the evidence behind the original choice, which allowed participants to partly resist feedback that was in fact false (Chapter 3). And in value-guided action, the exchange reached the representations that direct behavior: what participants observed of their own action's outcome was incorporated into their reported goal, the more strongly the weaker their value-based expectations had become (Chapter 4).

Across these very different paradigms, one observation recurs: revision was not overwriting. In none of the studies did external evidence simply replace the internal state it addressed. In Chapter 2, the incoming stimulus did not displace the memory of the episode; the two were held

in relation over time, and the relation itself carried information for the decision. In Chapter 3, a contradicting verdict did not directly move the confidence report by a fixed amount; it reopened access to stored evidence, and what that evidence returned determined how far the belief moved — a conclusion forced by the finding that no rule operating on the report alone could reproduce the behavior. In Chapter 4, the observed outcome entered a graded inference alongside internal signals, shifting the goal report in proportion to the balance of reliabilities rather than capturing it outright. In each case, the internal term retained standing in its own revision: the past was consulted, not merely corrected. This is what the iterative framing of Chapter 1 predicts, and it is the point on which the three studies, for all their differences, agree.

A second recurring observation is that the exchange is regulated, and regulated by signals that track reliability. In Chapter 2, the strength of the ongoing memory–input correspondence modulated how emerging perceptual information was translated into decision-relevant evidence — a signal that drew relatively heavily on frontoparietal regions, consistent with prefrontal and parietal cortex tracking the reliability of the evidence on which the decision was being built. In Chapter 3 the regulation was explicit and quantifiable: confidence determined whether the additional cycle of re-examination ran at all, and the threshold at which resistance appeared corresponded to the level at which a contradiction became more likely false than true given the feedback's error rate. In Chapter 4 the regulation appeared as reliability-weighting of the sources themselves: external evidence gained influence precisely when the internal prior had been weakened by volatility. At the same time, the studies counsel against picturing this regulation as finely tuned. In Chapter 3, participants' explicit estimates tracked the feedback source's drifting reliability, yet their trial-by-trial revisions barely used that knowledge; in Chapter 4, the shift toward external evidence under volatility exceeded what the fitted rational model prescribed. The

signals that govern the loop appear coarse and conservative — sufficient to protect the balance of persistence and flexibility in broad strokes, but far from the moment-to-moment optimality an ideal analysis would demand.

These are offered as themes, not as a demonstration of a single mechanism. The studies differ in domain, method, and level of analysis — electrophysiological characterization of perception, behavioral experimentation with a generative model of feedback integration, hierarchical Bayesian modeling of self-inference — and nothing in this dissertation shows that the iterations they describe are implemented by one process. Much remains particular to each: the network-level description of Chapter 2 has no counterpart in the other studies; the confidence-gated mechanism of Chapter 3 was identified behaviorally, without neural evidence; the central internal variable of Chapter 4, the motor intention, was inferred rather than measured. Whether the fast loop of perception, the post-decisional loop of confidence revision, and the slower loop of goal and value revision are nested instances of a common computation, or merely a family of resemblances, is a question this dissertation raises rather than settles.

That question does, however, suggest concrete work. The most direct next steps target the loops' hidden variables: an eye-tracking version of the goal-inference task, in preparation, converts the unobserved intention into a measurable quantity, allowing conflict between action and outcome to be defined independently of the report. A natural neural extension asks whether the re-examination inferred in Chapter 3 engages machinery of the kind observed in Chapter 2 — whether post-feedback revision recruits, in time-resolved recordings, a renewed comparison against maintained evidence; the confidence ratings already collected in the paradigm of Chapter 2 offer a first point of contact. Manipulation-based tests are equally available: if the resistance to

false feedback depends on consulting a stored trace, then loading memory between choice and feedback, or lengthening that interval, should selectively weaken it while leaving average feedback assimilation intact. And at the computational level, the three paradigms could in principle be addressed by models sharing an iterative core with task-specific components — an exercise whose value would lie as much in where such a model fails to generalize as in where it succeeds.

The dissertation's conclusion is correspondingly modest. In three settings spanning a glimpse, a judgment, and a goal, internal representations and external evidence met not once but repeatedly; revision proceeded by returning to what was held rather than by discarding it; and signals reflecting the reliability of each side helped decide how every round of the exchange was settled. Persistence and flexibility, on this evidence, are not competing dispositions between which a mind must choose, but the joint product of an iterative process that metacognition helps steer — a process by which a cognitive system, in an environment that does not stand still, decides what to keep believing by consulting both the world and itself, as many times as the question requires.