Metacognition as modal cognition

Kevin O'Neill, **David Lagnado**, and **Stephen M. Fleming**
{kevin.o'neill, d.lagnado, stephen.fleming}@ucl.ac.uk
Department of Experimental Psychology
University College London
26 Bedford Way
London WC1H 0AP

Abstract

An influential perspective is that metacognitive judgments involve forming propositional confidence in a self-centered frame of reference, evaluating the plausibility of propositions regarding one's own cognition and mental states. Here we build on this framework to propose that, because metacognition involves the consideration of alternative possibilities or hypotheses, it is an instance of a more general capacity known as modal cognition. By extension, using the Pearl causal hierarchy we distinguish between metacognition targeting conditional, interventional, and counterfactual probabilities, each of which allows one to make different kinds of inferences about one's own cognition. This view stresses the relevance of research on modal cognition for metacognition, highlights underexplored targets of metacognition, and helps explain differences between metacognitive phenomena.

Keywords: metacognition; modal cognition; confidence

Introduction

One of the most impressive, intriguing, and adaptive capacities of the human mind is metacognition: the introspection, evaluation, and control of our own cognition and mental states. An emerging perspective is that metacognitive judgments involve computing the subjective probability of a proposition from a self-centered reference frame (Fleming, 2024). That is, metacognition evaluates the plausibility of propositions regarding a self-model of one's own cognition (i.e., metarepresentation; Carruthers, 2009; Nelson & Narens, 1990).

However, this framework is agnostic as to how metacognitive computations are achieved. Here we develop a novel conceptual analysis of metacognition based on the insight that evaluating the plausibility of a proposition involves *modal cognition*—the consideration of alternative possibilities (Phillips & Kratzer, 2024). For example, when rating one's confidence that a stimulus was present, it is pertinent to ask questions such as "Would I have seen a stimulus if it were absent?" and "Could I perceive the stimulus if it were presented again?" (see, e.g., Mazor et al., 2025; Miyamoto et al., 2023). Under this perspective, metacognition evaluates modal queries over a self-model to determine the robustness of propositions about one's own cognition (Figure 1; Woodward, 2006, 2021). Here we distinguish three kinds of queries relevant to metacognition as defined by the Pearl causal hierarchy (PCH; Table 1; Bareinboim et al., 2022; Pearl, 2009).

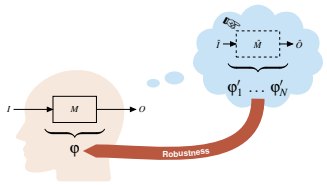

Figure 1: A schematic overview of metacognition. The agent considers a proposition φ that describes a cognitive mechanism M producing some output(s) O based on input(s) I. To determine the robustness of φ, the agent interrogates an approximate self-model $\hat{M}$ with questions about alternative possibilities (φ'_1 through φ'_N) that inform the robustness of φ. The pointing hand indicates that the agent may intervene on $\hat{M}$.

The causal hierarchy

The PCH distinguishes three kinds of information about a system: conditional, interventional, and counterfactual probabilities (Pearl, 2009). Critically, higher levels are more expressive than lower levels; as one climbs the PCH, additional representational and computational mechanisms are required (Bareinboim et al., 2022). To illustrate the relevance of the PCH for metacognition, we consider a signal detection task (Figure 2). Following the presence or absence of stimulus (S), the subject encodes a signal (X) which is susceptible to sensory noise (ε). Finally, the subject selects an action A by reporting "present" ($A = 1$) whenever $X > 0$, and "absent" ($A = -1$) otherwise.

Associations At the bottom level of the PCH, we can use conditional probabilities to identify associations between variables within this graph. For example, one can query the likelihood that one would report a present stimulus as being present, $P(A = 1|S = 1)$. Since $A = 1$ whenever $X > 0$, this reduces to $P(X > 0|S = 1)$, given by the normal CDF:

$$P(A = 1|S = 1) = P(X > 0|S = 1) = 1 - \Phi(-1) \approx .84$$

As another example, consider the posterior probability $P(S = -1|X = -2.5)$, representing the probability that the stimulus is absent given a signal $X = -2.5$. The posterior can be estimated using Bayes' rule, which (assuming flat priors over S) reduces to computing the relative likelihood of $X = -2.5$:

$$\begin{aligned} &P(S = -1|X = -2.5) \\ &\quad = \frac{p(X = -2.5|S = -1)}{p(X = -2.5|S = -1) + p(X = -2.5|S = 1)} \\ &\quad \approx .99 \end{aligned}$$

	Level	Typical Activities	Typical Question
3.	Counterfactual $P(Y_{X=x'} = y' \mid X = x, Y = y)$	Imagining, Explaining	Would I have seen a stimulus if one were present?
2.	Intervention $P(Y = y \mid do(X = x))$	Doing	Will I make the same choice under similar circumstances?
1.	Association $P(Y = y \mid X = x)$	Seeing, Predicting	What does my percept tell me about the stimulus?

Table 1: The PCH, adapted from Bareinboim et al. (2022).

Interventions At the next level up in the PCH, we can ask whether intervening on a variable would produce an outcome (the hand in Figure 1). Under the intervention $do(X = x)$, the subject severs causal pathways directed to X, forcing it to the value x independent of its usual causes (Figure 2, red X's). Unlike conditional probabilities, interventional probabilities carry information about causality: since S causes X, the intervention $do(S = s)$ changes the distribution of X, while the intervention $do(X = x)$ does not affect S. When the causal structure is known, interventional probabilities can be estimated from conditional probabilities by adjusting for confounding. In particular, interventions on root nodes with no incoming arrows (e.g., S) reduce to conditional probabilities (e.g., $P(A = 1|do(S = 1)) = P(A = 1|S = 1) \approx .84$).

Counterfactuals Finally, at the uppermost level, we can use counterfactual probabilities to ask about alternative pasts. Consider a trial in which the stimulus is absent ($S = -1$) and the sensory noise is strong ($\varepsilon = -1.5$), resulting in the sensory signal $X = -2.5$ and a correct decision $A = -1$. We can ask: would the subject have reported a stimulus if it were present (i.e., $P(A_{S=1}|X = -2.5, A = -1)$)?

Counterfactual probabilities involve three steps. First, given the observations $X = -2.5$ and $A = -1$, the subject infers the unknown variables S and ε as above. Given $X = -2.5$, the stimulus was likely absent with moderate noise, $P(S = -1, \varepsilon = -1.5|X = -2.5, A = -1) \approx .99$, but could have been present with substantial noise, $P(S = 1, \varepsilon = -3.5|X = -2.5, A = -1) \approx .01$. Second, the subject performs the intervention $do(S = 1)$ within each of these possibilities. Finally, they use the resulting distribution to predict their counterfactual action:

$$
\begin{aligned}
&P(A_{S=1} = 1|X = -2.5, A = -1) \\
&\quad = \sum_{s,\varepsilon} P(A_{S=1}(S = s, \varepsilon = \varepsilon) = 1) \\
&\qquad\qquad P(S = s, \varepsilon = \varepsilon|X = -2.5, A = -1) \\
&\quad = P(A_{S=1}(S = -1, \varepsilon = -1.5) = 1) \\
&\qquad\qquad P(S = -1, \varepsilon = -1.5|X = -2.5, A = -1) \\
&\qquad + P(A_{S=1}(S = 1, \varepsilon = -3.5) = 1) \\
&\qquad\qquad P(S = 1, \varepsilon = -3.5|X = -2.5, A = -1) \\
&\quad \approx 0(.99) + 0(.01) = 0
\end{aligned}
$$

Notably, this result disagrees with the interventional probability $P(A = 1|do(S = 1)) \approx .84$ because, while interventions generalize over trials, counterfactuals asks about *this particular trial*. Since $X = -2.5$, whether the stimulus was actually present or not, the subject can infer that ε would have been strong enough to prevent detection of any present stimulus.

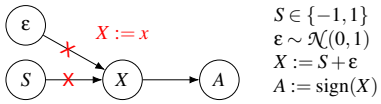

Figure 2: A causal graph for signal detection theory. Depending on the presence of a stimulus S, a signal X is encoded with noise ε. An action A based on the sign of X reports the detection of a stimulus. Red X's depict the intervention $do(X = x)$.

Assessing existing theories

Existing computational accounts of metacognition lie at the bottom of the PCH, since they define metacognition in terms of conditional probabilities (e.g., the probability that a decision is correct; Hangya et al., 2016; Pouget et al., 2016). But there is emerging evidence that metacognition might also involve interventions and counterfactuals. Though commonly formulated using conditional probabilities, choice consistency accounts of metacognition suggest that confidence tracks whether one would make the same choice if presented with the same problem again—a computation best described as an interventional probability (Boundy-Singer et al., 2023; Caziot & Mamassian, 2021; De Martino et al., 2013; Koriat, 2012). For instance, if stimulus presence is correlated with an external cue, computation of self-consistency should de-confound the stimulus and cue through intervention. Moreover, a recent model of metacognition about absence invokes counterfactuals: here, the agent determines whether they would have perceived a stimulus that could have been presented (Mazor et al., 2025). Unless inferences are the result of interventions restricted to the current trial, however, such models do not yet exploit the full expressivity of counterfactuals. More broadly, developing theories of metacognition within the framework of the PCH constitutes a promising avenue for future work.

Discussion

Recent accounts construe metacognition as estimating propositional confidence within a self-centered frame of reference (Fleming, 2024). Here we point out that under such accounts, metacognition entails the consideration of alternative possibilities, a faculty known as modal cognition (Phillips & Kratzer, 2024). Based on this observation, we suggest that modal considerations place computational and representational constraints on metacognition. Beyond the PCH, modal cognition is also sensitive to other features, including temporal orientation and specificity (Addis & Szpunar, 2024), plausibility (De Brigard & Parikh, 2019; Miceli et al., 2024; Morales-Torres et al., 2025), self-relevance (De Brigard et al., 2015; Khoudary et al., 2022), value (Bear et al., 2020; Morris et al., 2021; Phillips et al., 2019), controllability (McCloy & Byrne, 2000; Roese & Olson, 1995), and action/inaction differences (Byrne & McEleney, 2000). By considering metacognition in this framework, we aim to explain differences between metacognitive tasks, highlight understudied varieties of metacognition, and guide theory development in metacognition research.

Acknowledgments

SMF is a CIFAR Fellow in the Brain, Mind & Consciousness Program. This work was supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [selected as ERC Consolidator, grant number 101043666].

References

Addis, D. R., & Szpunar, K. K. (2024). Beyond the episodic–semantic continuum: the multidimensional model of mental representations. *Philosophical Transactions B*, *379*(1913), 20230408.

Bareinboim, E., Correa, J. D., Ibeling, D., & Icard, T. (2022). On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of judea pearl* (pp. 507–556).

Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., & Cushman, F. (2020). What comes to mind? *Cognition*, *194*, 104057.

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

Byrne, R. M., & McEleney, A. (2000). Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1318.

Carruthers, P. (2009). Mindreading underlies metacognition. *Behavioral and brain sciences*, *32*(2), 164–182.

Caziot, B., & Mamassian, P. (2021). Perceptual confidence judgments reflect self-consistency. *Journal of Vision*, *21*(12), 8–8.

De Brigard, F., & Parikh, N. (2019). Episodic counterfactual thinking. *Current Directions in Psychological Science*, *28*(1), 59–66.

De Brigard, F., Spreng, R. N., Mitchell, J. P., & Schacter, D. L. (2015). Neural activity associated with self, other, and object-based counterfactual thinking. *Neuroimage*, *109*, 12–26.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature neuroscience*, *16*(1), 105–110.

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

Hangya, B., Sanders, J. I., & Kepecs, A. (2016). A mathematical framework for statistical decision confidence. *Neural Computation*, *28*(9), 1840–1858.

Khoudary, A., O'Neill, K., Faul, L., Murray, S., Smallman, R., & De Brigard, F. (2022). Neural differences between internal and external episodic counterfactual thoughts. *Philosophical Transactions of the Royal Society B*, *377*(1866), 20210337.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological review*, *119*(1), 80.

Mazor, M., Moran, R., & Press, C. (2025). Beliefs about perception shape perceptual inference: An ideal observer model of detection. *Psychological Review*.

McCloy, R., & Byrne, R. M. (2000). Counterfactual thinking about controllable events. *Memory & Cognition*, *28*, 1071–1078.

Miceli, K., Morales-Torres, R., Khoudary, A., Faul, L., Parikh, N., & De Brigard, F. (2024). Perceived plausibility modulates hippocampal activity in episodic counterfactual thinking. *Hippocampus*, *34*(1), 2–6.

Miyamoto, K., Rushworth, M. F., & Shea, N. (2023). Imagining the future self through thought experiments. *Trends in Cognitive Sciences*, *27*(5), 446–455.

Morales-Torres, R., Miceli, K., Huang, S., Szpunar, K., & De Brigard, F. (2025). Episodic details are better remembered in plausible relative to implausible counterfactual simulations. *Psychonomic Bulletin & Review*, 1–8.

Morris, A., Phillips, J., Huang, K., & Cushman, F. (2021). Generating options and choosing between them depend on distinct forms of value representation. *Psychological science*, *32*(11), 1731–1746.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), (Vol. 26, p. 125-173). Academic Press.

Pearl, J. (2009). *Causality*. Cambridge university press.

Phillips, J., & Kratzer, A. (2024). Decomposing modal thought. *Psychological Review*, *131*(4), 966.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in cognitive sciences*, *23*(12), 1026–1040.

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

Roese, N. J., & Olson, J. M. (1995). Outcome controllability and counterfactual thinking. *Personality and Social Psychology Bulletin*, *21*(6), 620–628.

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*(1), 1–50.

Woodward, J. (2021). *Causation with a human face: Normative theory and descriptive psychology*. Oxford University Press.