Endogenous Model Uncertainty and Curiosity under Cognitive Limits

Jimmy Lee*

November 2026

Click here for the latest version

Abstract

How does a cognitively limited mind decide whether to take an interpretation seriously enough to recognize that investigating it could pay? Neuroscience suggests that people automatically classify situations through familiar mental models (e.g., a farmer attributing a neighbor's good harvest to the weather), entertaining alternatives when observations become difficult to reconcile with their initial interpretation. This paper treats such shifts in "mindsets" as a choice of how much model uncertainty to entertain. Each mental model pairs a prior with a view of how observable features combine to reveal the state. After an initial signal, a decision maker with a costless default and latent alternatives with less precise priors jointly chooses model weights and how many signals to acquire, and repeats this choice as evidence accumulates. The key discipline is an evidence constraint bounding exploration capacity: an alternative's weight is capped by its Bayesian posterior—a cognitive limit that responds to stimuli. Curiosity emerges as the value of resolving the model uncertainty entertained. Two results follow from the brain's evidence-based reasoning. Because unfamiliar interpretations leave more to learn, model weights and information acquisition are complements. With the evidence constraint, this produces curiosity failures: people may never investigate interpretations that could change their decisions. Because interpretations compete for evidentiary support, even an alternative she would never entertain can stifle investigation of a promising one—distinguishing the theory from other behavioral models. Sustaining curiosity therefore requires managing the space of possible interpretations, which I endogenize through creative and critical thinking as complementary forms of cognitive dimension reduction.

JEL codes: D81, D83, D91, O33.
Keywords: curiosity, bounded rationality, behavioral decision making, mental models, endogenous model uncertainty, evidence constraint, information acquisition, creative and critical thinking.

*Global Poverty Research Lab, Northwestern University. Email: jimmylee@u.northwestern.edu.

A mind is not a vessel to be filled, but a fire to be kindled.

— Plutarch, *On Listening to Lectures*

The efficiency of a human problem solver depends critically on his ability to find a representation of the problem that restricts search to a small subspace of the total problem space.

— Allen Newell and Herbert A. Simon, *Human Problem Solving* (1972)

1 Introduction

Is seeing a doctor a path to better health, or a risk of the mistreatment once inflicted on people of one's race (Alsan and Wanamaker, 2018)? Is a neighbor's good harvest the result of the new seeds he planted, or simply of the weather? Is school-based agricultural education in rural Africa a way to build students' independent farming skills and scientific capabilities while spreading improved techniques to their households, or does it simply train students as farm labor at the expense of their studies, as many school gardens do (Lee, 2026)?

In each of these cases, how a person classifies the situation shapes her behavior decisively. Each classification tells her not only what to expect but where to look: a farmer who credits the weather has no reason to try the seeds herself. Importantly, when past experience leads a person to dismiss a promising opportunity, providing facts has limited effect, because whether her reading is correct becomes clear only through personal, repeated observations: repeated visits to the doctor, comparisons across plots and seasons, or visiting school farms and assigning trial plots to students, which few rural parents do regularly. Gathering this evidence takes time and resources, keeping track of an unfamiliar reading takes mental effort, and when past experience makes people confident in their familiar interpretation, such evidence promises too little to justify these costs. The focal question thus becomes how to stimulate curiosity—the drive for people to seek, on their own, the information required to explore new perspectives.

In fact, these examples reflect a general feature of cognition. A large body of neuroscience, most prominently cognitive map theory (O'Keefe and Dostrovsky, 1971; O'Keefe and Nadel, 1978; Hafting et al., 2005), suggests that people do not weigh every interpretation of a situation afresh. To economize on mental effort under cognitive limits, the brain instead assigns each situation to a familiar mental model[1] retrieved from memory of similar past experience, before deliberation begins (Behrens et al., 2018; Sanders et al., 2020). As observations become harder to reconcile with that model, it increasingly entertains alternative mental models, but there is also room for choice: the brain tends to favor mental models that promise higher payoffs if true and ignore those that are plausible but less rewarding (Shenhav et al., 2013; Lieder and Griffiths, 2020).

Allowing people to seek information and reason beyond what associative memory supplies, this paper models these contextual shifts in "mindsets" as a choice of how much model uncertainty to entertain after an initial signal—such as how a development program is presented. These "mindsets", which are competing models of the world formally known as mental models, are defined as a prior over the payoff-relevant state together with a likelihood that says how observable features combine to reveal that state. The likelihood is what keeps mental models apart: interpretations that differed only in their priors over the same return could be combined into a single prior, whereas each mental model determines what counts as evidence for it. Two mental models therefore differ not only in what they expect but in what they attend to and how they read

[1] I use "mental models" and "cognitive maps" interchangeably in the limited sense of structured representations of which observable features matter and how they reveal an unobserved state. Cognitive map theory originated in studies of spatial navigation (Tolman, 1948; O'Keefe and Nadel, 1978) and has since been extended to abstract knowledge (Behrens et al., 2018). The paper formalizes this shared representational structure without making claims about neural implementation.

it: a farmer who credits the rainfall tracks the weather, and one who credits the seed tracks the plots. What each attends to is not chosen feature by feature (as in models of selective attention, e.g., Schwartzstein, 2014; Hanna et al., 2014) but follows from the explanation each gives of how the state is produced. The decision maker is not restricted to holding only one mental model at a time, and weighting mental models changes how a person interprets a situation. The decision maker starts from a default mental model that is costless to use and faces a menu of latent alternatives whose priors are less precise than the default's and that yield expected payoffs that differ from those under the default. After observing the initial signal, she jointly chooses (i) the attention weight she places on each alternative and (ii) how many subsequent signals to acquire. She repeats this choice as evidence accumulates.[2]

Regarding the choice of mental model weights, the framework integrates two decision theories under cognitive limits. As in Bayesian model averaging (e.g., Brock, Durlauf, and West, 2003; Cogley and Sargent, 2005), the posterior disciplines the weights, but the decision maker chooses them rather than adopting them: tracking an unfamiliar reading takes mental effort, and holding it forces a single action to hedge across models that recommend different actions, so she may place less weight on an alternative than its posterior. Akin to models of motivated beliefs (e.g., Brunnermeier and Parker, 2005; Bénabou and Tirole, 2016), expected payoffs guide that choice. However, I impose a key discipline, motivated by neuroscience: because the brain activates a mental model according to how well it predicts what is observed relative to its rivals, an evidence constraint caps the weight on an alternative at the posterior probability that her own mental models' reading of the evidence assigns it, which bounds her capacity to explore.[3] She can therefore neglect evidence but cannot invent it. This matches the psychology of motivated reasoning, in which people do not believe whatever they wish but construct justifications a dispassionate observer could accept, so that strong enough disconfirming evidence moves their beliefs (Kunda, 1990; Redlawsk et al., 2010).

In summary, the decision maker places weight on an alternative to the extent that acting on it promises, if it is right, a higher expected payoff than acting on the default, net of the costs of holding it, subject to the evidence cap. Much like collateral constraints in models of poverty traps (e.g., Galor and Zeira, 1993; Banerjee and Newman, 1993), the cap generates threshold effects in the exploration of alternative mental models. Consider an alternative that is more promising than the default—less promising ones are never activated. If it is right and evidence kept arriving, the cap would drift upward and the weight would follow. Two features of the setup break this convergence: the initial signal fixes the evidence cap, and model weights and information are complements, since information is worth more under the vaguer alternative than under the precise default, so the more weight she places on the alternative, the more she seeks. Together they create a threshold. A weak initial signal caps the weight below the level at which investigating the alternative pays, so the decision maker never investigates it, the cap never rises, and the alternative stays unexplored—a curiosity failure in which she fails to investigate an interpretation that could change her decision. The friction therefore lies in how she reads the situation rather than in the price of information. She can afford the evidence but sees no reason to seek it, and so her ignorance of the alternative persists. A signal just above that level starts a self-reinforcing cycle in which the weight justifies information, the information raises the cap, and the higher cap justifies more weight, so small differences in the initial signal produce sharply different subsequent investigation, even when the cost of information is the same.

Characterizing the choice of model uncertainty allows curiosity to be defined formally, as an endogenous object rather than a fixed trait. Once an alternative carries weight, the decision maker must act without knowing which mental model is right, which forces her to hedge and makes resolving the difference valuable. Curiosity is defined as the value of resolving that model uncertainty, a function of model weights: zero under

[2]The mechanism survives a two-stage extension in which the decision maker also values information for the weight it may license in the future. Under her own beliefs the evidence cap cannot be expected to rise, so this extra incentive can tip only alternatives already close to activation (Proposition OA.3 in the Online Appendix).

[3]This does not mean that the decision maker can never update more strongly than a Bayesian would in the usual sense, since what counts as evidence depends on her mental models.

the default alone, greatest when the mental models are most balanced, and falling away as either prevails, matching the well-known finding that curiosity peaks at intermediate confidence (Loewenstein, 1994; Kang et al., 2009). The framework can also incorporate a non-instrumental motive for curiosity through ambiguity aversion. A concave aggregator over the weights she chooses makes alternatives harder to activate, with an additional cost that grows with the gap in expected payoffs between mental models. When resolving model uncertainty is valued for its own sake, this ambiguity premium makes learning which model is right valuable even if the answer would not change her action. High-stakes questions—such as whether fertilizer bought on credit will double a farm's income or, if it is fake, leave the family in debt (Bold et al., 2017; Michelson et al., 2021)—are therefore harder to open but, once open, more actively pursued.

Curiosity also manifests in the type of information the decision maker seeks, once her exposure to signals is made endogenous. The choice of signal trades off learning more within the mental models she holds (refinement) against learning which of them is right (resolution), on terms set by the weight she places on the alternative. With little model uncertainty she seeks few signals, and those the default reads; as model uncertainty grows, curiosity directs her toward the features on which the mental models disagree, most strongly when the answer would change what she does; once the alternative dominates, she seeks signals the alternative reads. Shifts in "mindsets" are thus reflected in what information the decision maker seeks: a farmer who credits the weather watches the weather; one who suspects the seeds tries them on a plot of her own; one who has adopted them studies how to grow them well. Making one interpretation more plausible than another therefore changes not only how much information she seeks but which. Because the value of each signal is set by the weights she places on her mental models, the reverse also holds: the signals a person seeks, and the amount she will pay for them, help reveal the mental models she holds. Section 3.6 discusses various ways in which records of what farmers ask digital advisory services or AI assistants can be used to observe these choices at scale, and so to infer how interventions shift the mental models behind them.

What, then, do we gain from modeling humans' capacity to be curious, on top of a default classification based on associative memory? My framework helps characterize conditions under which two hallmark behavioral effects of associative-memory models persist. The first is context effects: adding payoff-irrelevant cues changes which representation of the situation comes to mind and so changes behavior (Bordalo et al., 2026); a consumer sees a jar of jam in plain packaging as a routine purchase and focuses on its price, but the same jar in fancy packaging becomes a treat and draws her attention to its quality. In my framework, a cue that leaves the default unchanged is an observable feature of the initial signal and counts as evidence like any other: it shifts weight only toward a mental model that predicts it better than the others, so cues sway behavior only if they are payoff-relevant under some mental model.[4] When a cue activates such an alternative, the curiosity this triggers tests it: with unbiased evidence the shift tends to persist if the alternative fits better than the default and to reverse otherwise, and when one of the two mental models is true, a tendency to seek confirming evidence only slows the reversal. In the model, a persistent context effect therefore points to evidence that is scarce, costly, or biased, or to a menu with no true mental model, rather than to a fixed cognitive bias.

Thus far, the discussion has pitted a single alternative against the default, but practical questions often admit many interpretations: a farmer who sees a neighbor's good harvest may credit the weather, the soil, a new seed variety, fertilizer, or the timing of planting. My framework allows many alternatives: the decision maker holds every alternative but at most one either not at all or at its evidence cap. Because the brain's automatic matching of cues against stored mental models weighs all of them at once, and she cannot drop a rival merely because she would not act on it, the support the evidence lends one alternative shrinks with every rival that also fits it. Interference effects, the second hallmark, therefore also work differently. In associative-memory models, experiences compete for retrieval and the representation with the greatest cue similarity comes to mind (Bordalo et al., 2025). Here, rivals that fit the same evidence divide its support, so several alternatives that each outperform the default can all fail to be activated, and even a cue that lends plausibility

[4] I take as given which mental model associative memory retrieves as the default. A cue that changes the default itself acts through retrieval, which I leave to future research.

to an alternative the decision maker would never entertain can reduce the information she seeks about a promising one, though she seeks none about the rival. Because this effect scales with the rival's plausibility rather than its appeal or the attention weight it receives, it distinguishes the framework from accounts based on retrieval or limited attention.[5] Adding interpretations to the menu thus cuts both ways: a first alternative can start curiosity, but enough rivals that fit the same evidence can extinguish it.

Interference effects thus raise a question that a fixed menu of mental models cannot answer: how does a mind with limited capacity keep its menu small enough to explore? I next endogenize the menu and introduce creative and critical thinking as two complementary forms of cognitive dimension reduction. Both rest on a premise from neuroscience: mental models are assembled from reusable pieces of explanation—formally known as templates—each a causal mechanism linking an underlying cause to observable features, such as rain to wet soil and high yields (Behrens et al., 2018; Whittington et al., 2020), and the brain imagines new situations by recombining stored experience (Hassabis et al., 2007; Schacter and Addis, 2007). Formally, I represent templates as latent factors in a linear Gaussian factor model: each factor's loading records which observable features the cause moves, and a mental model is a set of templates whose factors jointly generate what is observed.

As with curiosity, creative thinking is evidence-based: the decision maker builds new mental models from templates she already holds, and a new one is admitted only if the evidence favors it over the one it would replace. Vertical creative thinking adds a template to an existing mental model, explaining the same situation more deeply, such as crediting a neighbor's yield to the soil as well as the rain; it is admitted when what the current mental model fails to predict is surprising enough along the new template's direction. Horizontal creative thinking extends a mental model to new situations, such as carrying the arithmetic that children use in market transactions over to the academic problems taught in school, which does not happen automatically (Banerjee et al., 2025); it is admitted when the situations depart from expectations in the same direction. I prove that the two are not independent, yet neither substitutes for the other: a template that explains why a technique works sets how far a mental model can transfer, since reach is fixed by the templates it contains, but each transfer remains a horizontal step the decision maker must take and the evidence must justify, and such a template is worth nothing in a single situation, paying only through transfer. The same invariance that lets evidence pool across situations exposes the template to refutation when they disagree: the capacity to generalize and the exposure to refutation are two sides of one explanatory structure. Creative thinking is thus not a fixed trait but bounded by the templates the decision maker holds, so providing templates expands what she can create and how far it transfers. For training teachers and agricultural extension officers, this qualifies a common presumption that teaching why a practice works is enough to make trainees generalize: providing a generalizable explanation is not the same as developing the ability to generalize. Each transfer is a creative step the trainee must take and test in the new classroom or farm, and it happens only if her curiosity clears the threshold there; training in mechanisms therefore pays through her continued curiosity and horizontal creative thinking, not in place of them.

Creative thinking has a potential pitfall, however: a mental model assembled after seeing the evidence can fit it by chance. Critical thinking, a standing habit of mind, guards against this: it means not using the same data twice, so that every mental model, new or old, is judged only by evidence not used to form it. I model one core component of it, the discipline against after-the-fact fitting, as a parameter $\kappa \in [0,1]$: how strongly the decision maker discounts an explanation found after the fact for the breadth of the search that produced it, as in multiple hypothesis testing. I prove that a threshold value of κ, strictly below one and rising with the templates' vagueness, separates two regimes: below it, the probability of a false insight converges to one as the number of candidates searched grows, so discounting cannot prevent false insights in a broad search; above it, that probability converges to zero. Critical thinking has its own potential pitfall, and it is a curiosity failure rather than a rejection: the discount can hold a correct interpretation below the threshold at

[5] Section 3.7 shows that this observable implication also distinguishes the framework from standard Bayesian decision theories augmented with costly information acquisition, in which the rival is held at its posterior and so moves the decision maker's act.

which the decision maker would seek the evidence bearing on it, and more so the broader the search that found it.[6] Without complementary triggers for creative thinking, school-based agricultural education programs that cultivate critical thinking face a tradeoff between fewer false insights and slower adoption of correct ones: students trained to think critically should be less likely to adopt a new planting date or seed variety that merely coincided with a good harvest, but slower to adopt effective crop spacing they discover themselves.

Taken together, the results suggest that open-mindedness, curiosity, creative thinking, and critical thinking form a capability growth cycle within the brain's evidence-based reasoning. Open-mindedness—entertaining an alternative—starts curiosity, but curiosity tends to extinguish itself as the evidence it gathers resolves the uncertainty that prompted it. Creative thinking can replenish that uncertainty, since the surprises curiosity uncovers justify new mental models, while critical thinking judges each against fresh evidence and keeps chance fits off the menu, where they would dilute the evidence cap of the right alternative. None of the four is unambiguously beneficial on its own. Their value lies in how they interact, and cognitive dimension reduction is what makes the interaction manageable. A mind develops, on this view, not by collecting more information or entertaining more ideas, but by learning to generate interpretations worth investigating, to direct evidence toward them, and to eliminate those the evidence does not support. Throughout this paper, I illustrate the theory with development programs, teacher training, and agricultural technology adoption.

Related literature. This paper contributes to four strands of literature. First, I contribute to decision theory under model uncertainty. Existing theories take the weights on competing models of the world as given: in Bayesian model averaging, they are posterior probabilities;[7] in robust control, the decision maker guards against the worst case among models close to her benchmark;[8] and in smooth ambiguity preferences, the weights are part of her preferences.[9] Models of paradigm shifts let her switch models when the data fit her current one poorly enough, but the switch follows a statistical rule, and she holds one model at a time (Ortoleva, 2012; Ba, 2026). I turn the amount of model uncertainty a person entertains into an economic decision guided by payoffs, akin to models of motivated beliefs (Brunnermeier and Parker, 2005; Bénabou and Tirole, 2016), but disciplined by an evidence constraint. Abrupt paradigm shifts then become a special case: the decision maker switches abruptly to an alternative only if the models disagree sharply relative to the cost of attention weights, and otherwise adjusts her weight gradually. The two patterns parallel the gradual and abrupt changes neuroscience observes in how the brain represents an environment as it is morphed into another (Leutgeb et al., 2005; Wills et al., 2005; Colgin et al., 2008). As I model neural regularities in terms of Bayesian inference, the framework departs minimally from the Bayesian expected-utility benchmark: the decision maker updates by Bayes' rule within each model and maximizes expected utility given her weights; only the weights are chosen, under the evidence constraint. It also builds in a premise from neuroscience: interpretations, however novel, are assembled within the brain from what it already holds rather than supplied from outside (Dragoi and Tonegawa, 2011; Behrens et al., 2018). Unlike theories of misspecified learning (Esponda and Pouzo, 2016; Heidhues et al., 2018) or of persuasion with models (Schwartzstein and Sunderam, 2021), the question is therefore not whether the decision maker has access to the true model, but when evidence leads her to activate or create it. Last, the framework offers a theory for the growing evidence that people reason with different mental models and narratives (e.g., Shiller, 2017; Stantcheva, 2021; Andre et al., 2022): the signals a person seeks and the amount she will pay for them help reveal the mental models she holds.

Second, I add to the economics of information. From the value of information (Blackwell, 1953) through

[6] The pitfall is not an artifact of modeling only the discipline against after-the-fact fitting: any discipline that lowers the weight on an untested interpretation lowers the evidence sought about it, so testing a discounted insight requires curiosity about hypotheses one does not yet hold, or a commitment made before the evidence.

[7] See, among others, Brock, Durlauf, and West (2003) and Cogley and Sargent (2005).

[8] See Hansen and Sargent (2008).

[9] See Klibanoff et al. (2005, 2009), and Marinacci (2015) for a review of decision making under model uncertainty.

sequential search and experimentation to rational inattention, the decision maker knows what she is uncertain about and chooses how much to learn.[10] In bandit and experimentation models, how seriously she takes each option is fixed by her prior, so she always recognizes the option value of experimenting and declines only when it falls short of its cost. Failures to learn therefore arise only from costs, misspecified priors, unattended variables (Schwartzstein, 2014; Hanna et al., 2014), or unawareness of hypotheses that only an exogenous discovery can reveal (Karni and Vierø, 2013). Here, the decision maker has the same capacity to explore, but it is subject to a cognitive limit that itself responds to stimuli: the demand for information depends on which interpretations she entertains and on how much weight the evidence allows her to place on them, so evidence can be cheap, available, and decisive and still never be sought. This gives a choice-theoretic basis to the psychological account of curiosity as a response to an information gap that opens once a question is posed (Loewenstein, 1994; Golman and Loewenstein, 2018). Curiosity failures thus differ from existing sources of ignorance: the evidence is not prohibitively costly, avoided (Golman et al., 2017), or unimagined; it bears on an alternative mental model that carries too little weight to make it worth seeking.

Third, I add to behavioral models of memory and categorization. Associative-memory models explain choices by what comes to mind: payoff-irrelevant cues change choices (context effects), and similar experiences crowd one another out of memory (interference effects).[11] Coarse thinking likewise groups situations into categories and carries the content of information from one to another (Mullainathan et al., 2008). In both, classification feeds directly into the decision. I add a deliberation stage in between, in which the decision maker weighs the mental models retrieved and seeks information, disciplined by the evidence. The framework thus captures both people's capacity to correct themselves and the cognitive limits that hinder it. Methodologically, this paper joins work that derives restrictions on choice from neural regularities,[12] here through the evidence constraint, without requiring the validation that Gul and Pesendorfer (2008) argue neural data cannot provide.

Finally, I contribute to the economics of skills and capabilities. The human capital tradition holds that education's deepest return is adaptive: it raises the ability to deal with disequilibria (Schultz, 1975), speeds the adoption of new technologies (Nelson and Phelps, 1966), and pays more where technical change is fast (Foster and Rosenzweig, 1996). The skill-formation literature models capabilities as stocks produced by investment, with self-productivity (skills beget skills) and dynamic complementarity (skills raise the return to later investment).[13] Yet general thinking skills transfer poorly across domains (Willingham, 2007; Sala and Gobet, 2017), and while mechanistic explanations help farmers generalize (Sankar et al., 2026), conceptual knowledge is modeled as a known truth valuable within a single decision (Davies and Sankar, 2026), which leaves unexplained why knowing why something works transfers to some new situations but not others. I model curiosity, creative thinking, and critical thinking not as stocks carried across settings but as processes set in motion by the evidence of particular situations. What transfers is a library of templates and the discipline of judging interpretations on fresh evidence; each extension to a new situation is a step the evidence must justify. Knowing why something works thus makes transfer possible but not automatic, which reconciles the two findings.

Organization. Section 2 introduces the model through a simple example and then presents the general setup. Section 3 presents the main results. Section 4 characterizes the informational environments under which hallmark behavioral effects in associative-memory models persist. Sections 5 and 6 endogenize the menu through creative and critical thinking. Section 7 concludes.

[10] See, among others, Gittins (1979), Weitzman (1979), Sims (2003), Matějka and McKay (2015), Caplin and Dean (2015), and Maćkowiak et al. (2023) for a review.

[11] See, among others, Gennaioli and Shleifer (2010), Bordalo et al. (2020, 2023, 2025), Enke et al. (2024), and Bordalo et al. (2026).

[12] See, for instance, Woodford (2012), Khaw et al. (2021), and Webb et al. (2021) on efficient coding and normalization in valuation, and Bernheim (2009) on the case for this approach.

[13] See, among others, Heckman et al. (2006), Cunha and Heckman (2007), Borghans et al. (2008), Cunha et al. (2010), Almlund et al. (2011), Heckman and Kautz (2012), and Heckman and Corbin (2016), who relate skills to Sen's capabilities.

2 Setup

The model is inspired by cognitive map theory, which originally dealt with spatial navigation but has recently been extended to learning about conceptual ideas, memory mapping, and social relationships (Constantinescu et al., 2016; Behrens et al., 2018; Whittington et al., 2020). A decision maker (DM) holds a familiar default interpretation of her situation, which runs costlessly as an "auto-pilot" mode (Kool et al., 2010; Shenhav et al., 2013), and latent alternatives, which take cognitive effort to hold (Shenhav et al., 2017; Lieder and Griffiths, 2020). After an initial observation, she chooses how much attention weight to place on the alternatives and how much costly information to acquire, subject to an evidence constraint, and then takes a single act. I first convey the main intuitions through a simple example, in which each interpretation is just a prior, and then present the general setup.

2.1 An illustrative example of mindset shifts

The setting. A farmer sees that a neighbor who planted a new seed variety harvested far more than usual. She must decide how extensively to adopt the variety. Before committing land to it, she can learn more in several ways: asking farmers who have grown it how it did, visiting their fields, consulting an extension agent, or trying a small patch herself. Let $\theta \in \mathbb{R}$ denote the unknown return to adopting the variety on her land. She chooses an adoption scale a and earns $u(a,\theta) = a\theta - a^2/2$: each unit adopted returns θ, and a larger scale is increasingly costly to manage. With an expected return m she chooses $a = m$, assuming an interior solution: the higher the estimate, the more extensively she adopts.

Two interpretations. The farmer, the DM, holds two interpretations of the neighbor's good harvest, here simply two priors over θ:

- the *default* ($k = 0$), "the neighbor's harvest was probably due to good weather, so the variety is unlikely to do much better on my land": $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$, with a modest mean and a high precision $\tau_0 \equiv 1/\sigma_0^2$, because experience has made her confident that new varieties rarely make much difference;
- an *alternative* ($k = 1$), "the variety made the difference and may substantially raise yields on my land": $\theta \sim \mathcal{N}(\mu_1, \sigma_1^2)$, with $\mu_1 > \mu_0$ but $\tau_1 < \tau_0$, because this reading is more promising but unfamiliar.

Signals and choices. The DM first observes how much more the neighbor harvested than usual, an initial signal $x = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, v^2)$, where the noise absorbs the neighbor's weather, soil, and management; a demonstration plot can provide such evidence. Each interpretation reads x by Bayes' rule, giving an estimate m_k of the return. Because the default is precise, its estimate barely moves with x: read through the default, a big harvest is absorbed as good weather. The alternative's estimate moves much more. The DM then chooses an attention weight $w \in [0,1]$ on the alternative, with $1-w$ on the default, and a number n of inquiries, in any of these forms, at cost c each in time, travel, favors, or seed, each bringing back an independent noisy report on the variety, $y_i = \theta + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, v_y^2)$. She then chooses her adoption scale.

The objective. Let $v_k := m_k^2/2$ be what acting on interpretation k is worth if it is right, $g_k(n)$ what n inquiries would add to that value by refining its estimate, and $D := (m_1 - m_0)^2/2$ the disagreement between the two interpretations' recommended acts. The DM maximizes

$$V(w,n) \;=\; \underbrace{(1-w)\,v_0 + w\,v_1}_{\text{value of acting}} \;-\; \underbrace{w(1-w)\,D}_{\text{hedging cost}} \;+\; \underbrace{(1-w)\,g_0(n) + w\,g_1(n)}_{\text{value of inquiries}} \;-\; \tfrac{\gamma}{2}w^2 \;-\; cn.$$

Because she takes one act for both interpretations, any compromise is wrong under both, and the *hedging cost* is the resulting loss. It is zero when she holds one interpretation and largest at balanced weights, and

it is exactly what she would gain by learning which interpretation is right. The term $\frac{\gamma}{2}w^2$ is the mental effort of keeping the unfamiliar reading in mind. The objective separately credits the value of refining each interpretation's recommendation, evaluated at the weights chosen before the information arrives. Because the alternative is vaguer, it has more to learn: $g_1(n) > g_0(n)$.

The evidence constraint. The weight is a choice guided by payoffs, but it is capped by the evidence:

$$w \leq \overline{W}_1(x) = \frac{L_1(x)}{L_0(x)+L_1(x)},$$

where $L_k(x)$ is how likely the signal is under interpretation k. The cap is the posterior probability that a Bayesian starting from even odds would assign to the alternative. The DM may place less weight on it than the evidence supports, but never more.

A curiosity failure. The value of inquiries rises with w because the alternative has more to learn: weight and information are complements. Suppose the default alone would not justify an inquiry, $g_0(1) < c$. Since each further inquiry teaches less than the first, she makes any only if

$$w > \underline{w} := \frac{c - g_0(1)}{g_1(1) - g_0(1)}.$$

Because $w \leq \overline{W}_1(x)$, a signal that leaves $\overline{W}_1(x) \leq \underline{w}$ means she never asks, however promising the alternative and even if it is right; and without new information the cap never moves. This is a *curiosity failure*: the DM fails to investigate an interpretation that could change her decision, not because information is unavailable, but because the evidence constraint keeps her weight on the alternative too low to make investigation worthwhile. A cap above $\underline{w}$ is necessary for inquiries but not sufficient, since she may choose a weight below the cap; in the example below, the cap binds over the relevant range. What she forgoes is investigation, not adoption: on the default's estimate she may still adopt the variety on a modest scale, but an ordinary harvest, which weather and soil also move, tells her little about the variety itself, so without seeking out such evidence the question of whether it deserves more stays unexamined.

Curiosity, the drive to seek the information required to explore new perspectives, is distinct from the decision to acquire information. It is the value of resolving the model uncertainty she entertains: the hedging cost $w(1-w)D$, what she would gain by learning which interpretation is right. It is absent while she holds the default alone and arises once she takes the alternative seriously, but leads her to seek information only if her weight clears the threshold.

Figure 1 illustrates a curiosity failure under specific parameter values. Over a wide range of harvests, the variety promises more than the weather reading, and she would take it more seriously than the evidence allows; yet the cap stays below the threshold and she never asks. Even a harvest exactly as large as the alternative predicts leaves the cap short of it. Only a strikingly good harvest lifts the cap above the threshold; from there, favorable reports raise the cap further, which justifies more weight and more inquiries.

Adding a rival. Now suppose the farmer also holds a second alternative ($k = 2$), "the neighbor's soil is more fertile than mine": the variety did well on fertile land and may do little on hers, $\theta \sim \mathcal{N}(\mu_2, \sigma_2^2)$ with $\mu_2 < \mu_0$. It recommends adopting on a smaller scale than the default, so it promises less and she never places weight on it. Yet her mind weighs it against the same evidence, so its likelihood enters the cap of the promising alternative:

$$\overline{W}_1(x) = \frac{L_1(x)}{L_0(x)+L_1(x)+L_2(x)}.$$

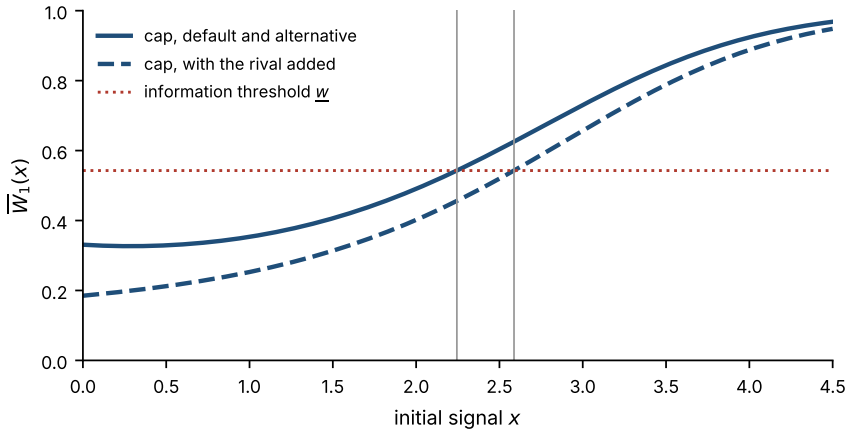


FIGURE 1. CURIOSITY FAILURE AND INTERFERENCE IN THE ILLUSTRATIVE EXAMPLE

Notes: The solid line is the evidence cap on the promising alternative with the default and that alternative on the menu; the dashed line adds the rival. The dotted line is the information threshold $\underline{w} \approx 0.54$; the DM makes inquiries only where the cap lies above it, from $x \approx 2.24$ without the rival and $x \approx 2.59$ with it. Each interpretation's estimate after x is $m_k = (\tau_k \mu_k + \nu^{-2} x)/P_k$ with $P_k = \tau_k + \nu^{-2}$, so $g_k(n) = [1/P_k - 1/(P_k + n\nu_y^{-2})]/2$, and $L_k(x)$ is the normal density with mean μ_k and variance $\sigma_k^2 + \nu^2$. Parameters: $\mu_0 = 1$, $\tau_0 = 4$; $\mu_1 = 2$, $\tau_1 = 0.5$; $\mu_2 = 0$, $\tau_2 = 1$; $\nu^2 = \nu_y^2 = 1$; $\gamma = 0.9$; $c = 0.08$.

In Figure 1 (dashed), a harvest that would have started her inquiries now leaves the cap below the threshold: she stops investigating the promising alternative, although she places no weight on the rival and seeks no information about it. Only a still better harvest starts her inquiries. This is an *interference effect*: a plausible interpretation the DM would never act on stifles her investigation of a promising one, through its likelihood rather than the weight it receives. The rival moves her act only indirectly, through the lower weight on the promising alternative. Under Bayesian model averaging, the rival would instead carry its posterior weight and pull her act toward its own recommendation; under limited attention, a rival she does not attend to would have no effect.

Take a harvest of $x = 2.5$. Facing only the default, the farmer investigates the promising interpretation; once the rival is added, she does not. Nothing about the harvest or the cost of inquiry has changed, and the rival receives no weight. The example thus illustrates two departures from conventional accounts of information acquisition: a promising interpretation may never be investigated, and even an interpretation the DM would never entertain can suppress investigation of another. Whether she learns depends not only on what she knows but on which interpretations her mind weighs against the evidence.

2.2 The general setup

The example explains why a farmer may fail to investigate an alternative interpretation, but leaves two questions about its interpretations open. First, if they differ only in their priors over the same return, a Bayesian decision maker could combine them into a single prior, so what makes them separate objects of attention? An interpretation does more than express an expectation: it specifies how observable features reveal the state, and so what counts as evidence for it. I therefore define a mental model as a prior paired with such a mapping. Second, why may the DM place less weight on an interpretation than its posterior, but not more? Because evidence is judged internally, by her own mental models rather than against an external standard, and her deliberation cannot overrule them. She can take an interpretation less seriously than the evidence supports, but she cannot make the evidence support it more. The posterior is therefore a ceiling, not a target.

The example also treats every inquiry as a noisy signal of the same return, leaving unexplained why different interpretations lead people to seek different information. Because mental models read different features,

shifts in the weights placed on them can change the directions of information acquisition. Mental models thus determine how observations are interpreted as evidence, attention weights determine how seriously each interpretation is entertained, and together they shape how much information the DM seeks and where she looks. The general setup also allows many competing interpretations and broader preferences and payoffs, enabling me to extend the definition of curiosity, characterize curiosity failures more generally, and derive results beyond the example.

2.2.1 Mental models, signals, and timing

Let $\theta \in \mathbb{R}$ denote an unknown payoff-relevant state: the return to a course of action whose value the DM does not know. The DM has von Neumann–Morgenstern preferences represented by $u(a,\theta)$ over acts $a \in \mathscr{A} \subseteq \mathbb{R}$ and states θ, and $\mathbb{E}_\mu u(a,\theta)$ denotes expected utility under a belief μ.

Initial signal. The DM observes an initial signal, a raw feature vector $x \in \mathbb{R}^d$. For example, what a farmer sees of her neighbor's harvest includes the yield, the rainfall that season, the variety planted, and how the plot was managed. The initial signal is exogenous to the DM but can be designed by an intervention: a demonstration plot or an extension message that highlights the variety would be one such x.

Mental models. The DM holds a finite menu of mental models of the situation she faces, indexed by $k \in \{0,1,\dots,K\}$. Index $k=0$ is the *default* model, which she has applied many times before and which she operates on costlessly; indices $k=1,\dots,K$ are *latent alternatives*, which she may or may not bring to bear.[14] The default is the mental model that associative memory retrieves for the situation, and I take that retrieval as given: a contextual cue may change which mental model serves as the default, but the analysis begins once the default is set, and a former default then joins the latent alternatives. The framework thus models the deliberation stage that follows retrieval; I leave modeling retrieval itself to future research. Until Sections 5 and 6, the menu is fixed: subsequent signals reweight and refine the models on it rather than introducing new ones. The menu is the DM's own. An outside party can target a mental model on it and activate it, by designing the initial signal, but cannot supply one.[15]

A mental model has two aspects. The first is a *prior* over the state: model k asserts $\theta \sim \mathscr{N}(\mu_k, \sigma_k^2)$ with precision $\tau_k \equiv 1/\sigma_k^2$, and the default is the most precise on the menu, $\tau_0 > \tau_k$ for every $k \geq 1$, because it has been applied most often. The means are left unordered. The second aspect is a *generative process* that says how the observable features combine to reveal the state. This is the model's explanation of how the situation produces what is observed, and it determines both which features the model attends to and how it interprets them. As cognitive map theory treats it, each model is a complete generative process for the features of the initial signal:

$$\theta \sim \mathscr{N}(\mu_k, \sigma_k^2), \qquad x \mid \theta \sim \mathscr{N}(\theta\, b_k,\, \Sigma), \qquad \tau_k := 1/\sigma_k^2, \qquad k = 0,1,\dots,K, \tag{1}$$

where $b_k \in \mathbb{R}^d \setminus \{0\}$ is the mental model's loading vector and $\Sigma \succ 0$ is a common noise covariance. Each alternative is structurally distinct from the default and from the other alternatives, in that it carries its own loading b_k.

[14] A latent alternative need not come from experience; it is available but carries no weight until activated. Latent structure of this kind has a neural counterpart: sequences that later form the mental model of a new environment are already present in the hippocampus before that environment is first experienced (Dragoi and Tonegawa, 2011).

[15] There is evidence that the brain comes with built-in templates from which it constructs novel mental models in response to environmental stimuli (Section 5). In rats, the hippocampal sequences that will represent a new environment are already present before the animal enters it; exposure selects and activates one of these preexisting maps rather than installing it (Dragoi and Tonegawa, 2011). The environment, like an intervention, determines which map is activated, not which maps exist.

Each mental model also implies an *attention mapping*: the weights it places on the components of the initial signal when it forms its belief about θ. Features a model does not encode are not learned from, however long the experience. Hanna et al. (2014) document such a failure among seaweed farmers who never learned from years of their own harvests that pod size mattered; in their model a farmer can keep track of only a few of the many inputs that affect yield. Here a farmer does not choose which inputs to track one by one: each mental model is an explanation of how yields are produced, and that explanation determines which features she looks at.

Lemma 1 (Attention mappings are derived)**.** *Under* (1)*, the posterior over θ given x depends on x only through the scalar*

$$s_k \;=\; \alpha_k' x, \qquad \alpha_k \;=\; \frac{\Sigma^{-1} b_k}{b_k' \Sigma^{-1} b_k}, \tag{2}$$

and $s_k \mid \theta \sim \mathcal{N}(\theta, \nu_k^2)$ with $\nu_k^2 = (b_k' \Sigma^{-1} b_k)^{-1}$. The posterior is Gaussian with precision $P_k(0) = \tau_k + \nu_k^{-2}$ and mean $m_k = \big(\tau_k \mu_k + \nu_k^{-2} s_k\big)/P_k(0)$.

Each mental model thus compresses the observable features into a single number s_k, and the weights in that compression, α_k, are its attention mapping: a consequence of the explanation the model gives, not a separate choice over features, and it changes when the model does. I assume throughout that the default remains the most precise model after reading the initial signal, $P_0(0) > P_k(0)$ for every $k \geq 1$, so that every alternative has more left to learn than the default.[16] When $b_j \neq b_k$ two models read the same raw observation differently, and features diagnostic under one may be ignored by the other, so shifting weight between them changes both what the DM looks at and how she interprets the same situation; when $b_j \propto b_k$ they attend identically and differ only in (μ_k, τ_k). Differences in priors alone, however, do not capture the distinct ways mental models interpret observable features.

Subsequent signals. After observing the initial signal x, the DM may acquire $n \geq 0$ subsequent signals at marginal cost c per signal, each of the form $y_i = \theta + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \nu_y^2)$.

Attention weights. The DM chooses a vector of attention weights $\mathbf{w} = (w_1, \ldots, w_K)$ on the alternatives,[17] with $w_k \geq 0$ and $\sum_{k \geq 1} w_k \leq 1$, the remainder $w_0 = 1 - \sum_{k \geq 1} w_k$ resting on the default; and she chooses the number of subsequent signals n. Timing is:

(1) The initial signal x is realized and observed.

(2) The DM chooses the weights w_k on the alternatives, each bounded above by an evidence constraint defined below, and the number of subsequent signals $n \in \{0, 1, 2, \ldots\}$.

(3) The n subsequent signals realize.

(4) The DM acts. The compromise across models is set at the weights chosen in step (2), and the signals refine each mental model's own recommendation.

The results concern the weights chosen at step (2). At step (4) the DM chooses a single act under model uncertainty, so the maximization runs over one a evaluated at the aggregated weights; aggregating model-contingent maxima would instead describe a DM who *knows* which mental model applies. Placing the maximum outside the sum over models is what gives the disagreement between mental models its bite, and it

[16] A situation has one clear default rather than several close contenders, as in cognitive map theory, where the brain settles on one well-learned map for a context (Wills et al., 2005; Sanders et al., 2020). Formally, one observation does not undo the advantage of repeated experience; the assumption holds whenever $\nu_k = \nu_0$.

[17] When the context is ambiguous, the brain flickers between familiar and new mental models and searches for evidence until it settles on one (Sanders et al., 2020): for fractions of a second in spatial navigation (Jezek et al., 2011), longer for ambiguous figures (Leopold and Logothetis, 1999), and plausibly months for the conceptual mental models studied here.

generates the hedging cost defined below. The model is thus a single choice followed by a Bayesian update, with the number of signals fixed before any is seen. Where the analysis follows the DM over several rounds of information, it repeats this stage, with the updated bounds carried forward and the DM choosing myopically in each.[18]

2.2.2 Preferences, costs, and the objective function

The DM evaluates an act by its expected utility under each mental model and aggregates across models with the weights. To nest attitudes toward model uncertainty, the aggregation is written through a strictly increasing $\varphi : \mathbb{R} \to \mathbb{R}$, as in the smooth ambiguity model (Klibanoff et al., 2005), except that the weights are chosen rather than given by preferences. The baseline sets $\varphi(U) = U$: the DM maximizes expected utility under the mixture of models and has no ambiguity aversion as such. A strictly concave φ would add ambiguity aversion, as discussed in Section 3.5.

Holding alternatives active imposes a cognitive cost,

$$C(\mathbf{w}) \;=\; \frac{\gamma}{2}\sum_{k\geq 1} w_k^2 \;+\; \eta \sum_{1\leq j<k\leq K} w_j w_k, \qquad \eta > \gamma > 0. \tag{3}$$

Operating under the default alone is costless; the cost is convex in each weight, since each additional unit of attention weight on an alternative requires progressively greater inhibition of the default and suppression of interference from other active alternatives;[19] and because $\eta > \gamma$ it is superadditive across alternatives: holding two at once costs strictly more than the sum of holding each alone. When a single alternative k is held, (3) is simply $(\gamma/2)w_k^2$.

Write $\mathbb{E}_k$ and $\mathbb{E}_{k,n}$ for expectations under mental model k's posterior before and after the n signals, and $U_k(x,n) := \mathbb{E}_y^k[\max_a \mathbb{E}_{k,n} u(a,\theta)]$, with $U_k(x) := U_k(x,0)$, for the value of acting on model k, the outer expectation taken over the signals as model k predicts them. The derived objective is

$$V(\mathbf{w},n;x) \;=\; \max_{a\in\mathscr{A}} \sum_{k=0}^{K} w_k\, \varphi\big(\mathbb{E}_k\, u(a,\theta)\big) \;+\; \sum_{k=0}^{K} w_k \big[\varphi\big(U_k(x,n)\big) - \varphi\big(U_k(x)\big)\big] \;-\; C(\mathbf{w}) \;-\; c\,n. \tag{4}$$

The first term is the single act chosen under the model uncertainty the weights express; the second credits each mental model with the refinement of its own recommendation. Within a stage, then, signals refine the mental models without settling which is right; that evidence takes effect when the weights are re-chosen.

2.2.3 The evidence constraint

The DM maximizes (4) subject to an evidence constraint, which limits her capacity to entertain alternative interpretations. The limit is set by the evidence as her mental models read it: the weight she can place on an alternative cannot exceed the support it receives relative to the competing mental models. Evidence thus sets the capacity to explore, and that capacity decides whether further evidence is ever acquired.[20] Let $L_k(x)$ be

[18] Myopia keeps each stage a static problem with explicit closed forms. It rules out buying signals today for the weight they would license tomorrow, the experimentation motive of bandit models; the neuroscience describes reassignment as ongoing inference rather than planned exploration (Sanders et al., 2020). The approach parallels intergenerational poverty-trap models, where dynamics arise from state variables carried across generations even when individual decisions are formulated over a limited horizon (Galor and Zeira, 1993; Banerjee and Newman, 1993; Mookherjee and Ray, 2003); here evidence plays the role of wealth. Little is lost: Proposition OA.3 in the Online Appendix shows that the main results survive a two-stage forward-looking extension.

[19] The convex form summarizes two neural constraints: working-memory capacity is limited to roughly three to four items (Cowan, 2001), and the precision with which items are held drops nonlinearly as load increases (Bays and Husain, 2008); tracking competing interpretations, each with its own predictions, is more demanding than tracking items.

[20] The constraint is motivated by neuroscience. Hippocampal remapping behaves like Bayesian inference over which context applies: the mental model that becomes active is the one favored by the posterior given the evidence (Sanders et al., 2020). Taking that

the marginal likelihood of the initial signal under mental model k and $\Lambda_k(x) := L_k(x)/L_0(x)$ the Bayes factor of alternative k against the default. Define

$$\overline{W}_k(x) \;:=\; \frac{L_k(x)}{L_0(x)+\sum_{j=1}^{K} L_j(x)} \;=\; \frac{\Lambda_k(x)}{1+\sum_{j=1}^{K} \Lambda_j(x)}, \qquad k=1,\ldots,K. \tag{5}$$

The right-hand side is the posterior probability that a Bayesian with a uniform prior over the $K+1$ mental models would assign to alternative k after seeing x, so the constraint $w_k \le \overline{W}_k(x)$ reads: *the DM may entertain an alternative less seriously than her mental models' reading of the evidence warrants, but not more*. When x is consistent with the default, every $\overline{W}_k \approx 0$; when x falls in the default's tails, the bounds of the alternatives that account for x rise.

The constraint differs from the cognitive cost of holding mental models. The cost prices how much model uncertainty she entertains in total; the constraint bounds, model by model, how seriously she can take each alternative, and it moves with what she observes. It is not a budget on total attention, $\sum_k w_k \le B$, but a separate ceiling for each alternative that the evidence raises or lowers.

Two limiting cases. The constraint places the DM between two decision theories. As in Bayesian model averaging (Brock, Durlauf, and West, 2003; Cogley and Sargent, 2005), the posterior disciplines the weights, but the DM chooses them rather than adopting them: were holding models costless, she would set every weight at its cap. Akin to models of motivated beliefs (Brunnermeier and Parker, 2005; Bénabou and Tirole, 2016), expected payoffs guide that choice: were there no cap, she would set the weights by the value of the models they favor. The constraint sits between the two because it is an inequality: she can leave an unpromising alternative unentertained, which a Bayesian cannot, while its evidence still counts against the others, which it would not without the cap. This discipline is consistent with what neuroscience finds in how the brain activates a mental model, according to how well it predicts what is observed relative to its rivals (Sanders et al., 2020; Wills et al., 2005; Jezek et al., 2011), and with what psychology finds in motivated reasoning: instead of unconstrained beliefs, people construct justifications that a dispassionate observer could accept, and strong enough disconfirming evidence moves their beliefs (Kunda, 1990; Redlawsk et al., 2010).[21] Neither establishes a hard cap on its own; I adopt the cap as an economic discipline, to be judged by its empirical content, discussed below.

Discussion of the evidence constraint. I make three remarks on the constraint. First, for evidence that favors an alternative, the ceiling can be derived rather than imposed. If deliberation chooses whether to entertain the alternative and how much of that evidence to take in, and the weight is the posterior given what is taken in, then every weight up to the cap is reachable and none above it (Lemma OA.1 in the Online Appendix). Attending selectively to favorable features is no way around this, since a different weighting of the features is a different mental model (Lemma 1) with its own cap. For evidence against an alternative, partial processing alone would lift the weight back toward the prior, so the constraint adds a restriction: what the automatic assignment has registered cannot be discarded. The DM can neglect evidence but cannot invent it. What the derivation takes as primitive is that deliberation sets how much weight the evidence receives, not what the evidence says, and the same primitive fixes the menu: which mental models are retrieved, and so which likelihoods enter the denominator of (5), is set by the automatic assignment before deliberation begins.

automatic assignment as $p_k(x) = \pi_k L_k(x)/\sum_j \pi_j L_j(x)$, deliberation can hold a model only to the extent the assignment has made it available, $w_k \le p_k(x)$; with a uniform π this is (5). The cap therefore depends on the menu, as judged probability does in support theory (Tversky and Koehler, 1994). A well-learned mental model is an attractor that snaps back after a brief disturbance (Wills et al., 2005; Jezek et al., 2011).

[21] Given the weights, the DM's belief about θ is the $\mathbf{w}$-mixture of the mental models' posteriors, a single distribution. The two-model language describes not what she believes but what governs it: a Bayes factor exists only between identified components, the costs are functions of the weights, and updating tracks which component each piece of evidence favors.

Deliberation can withhold attention weight from a retrieved model's evidence but cannot remove the model, which would raise the other caps above what the evidence justifies; a retrieved alternative leaves the menu when evidence against it arrives, not because the DM finds it unattractive. Second, the ceiling need not be hard. If weight above the posterior is available at a price per unit, a high enough price is never paid and every result is as stated; at lower prices the weight drifts above the posterior, continuously in the price, and only when exceeding it is free does the DM believe whatever pays, akin to unconstrained motivated beliefs (Proposition OA.1 in the Online Appendix). Third, the constraint has empirical content: a person's weight on an alternative should fall at or below the Bayesian posterior computed under her mental models from the evidence she was shown, never above it, and it should fall when rival alternatives are added to the menu.

2.2.4 Bayesian updating

Every signal triggers two types of Bayesian updating. Within each active mental model, the posterior over θ updates by Bayes' rule, which is automatic and refines precision. Across mental models, the same signal is evidence about which of them generated it. Write $\Lambda_k^y(y^n)$ for the likelihood ratio of the n subsequent signals under alternative k against the default, each mental model evaluated at its own step-(2) posterior. After the signals the weights and the evidence caps update by Bayes' rule,

$$w_{k,n} \;=\; \frac{w_k\,\Lambda_k^y(y^n)}{w_0+\sum_{j=1}^K w_j\,\Lambda_j^y(y^n)}, \qquad \overline{W}_{k,n} \;=\; \frac{\Lambda_k(x)\,\Lambda_k^y(y^n)}{1+\sum_{j=1}^K \Lambda_j(x)\,\Lambda_j^y(y^n)}, \tag{6}$$

and carry forward to the next stage. This is the recursive update of the second-order prior in Klibanoff et al. (2009), applied to weights the DM has chosen rather than inherited. Updating preserves the discipline: the odds of every alternative against the default never exceed its Bayes factor, and the weights are a martingale under the mixture, so acquiring signals cannot be expected to move attention weights in either direction, only to resolve it (Lemma OA.2 in the Online Appendix).[22]

2.2.5 The hedging cost of holding several mental models

A single act must serve whatever mixture of mental models the DM holds. The *hedging cost* of holding weights $\mathbf{w}$ is the value lost because she must choose one act rather than the act each mental model recommends:

$$A(\mathbf{w}) \;:=\; \underbrace{\sum_{k=0}^{K} w_k\,\varphi\big(U_k\big)}_{\text{value of knowing the mental model}} \;-\; \underbrace{\max_a \sum_{k=0}^{K} w_k\,\varphi\big(\mathbb{E}_k u(a,\theta)\big)}_{\text{value of one act under model uncertainty}}, \tag{7}$$

where $U_k = U_k(x) = \max_a \mathbb{E}_k u(a,\theta)$ is the model-contingent value. The hedging cost is nonnegative, zero when all weight rests on one mental model or when every model receiving weight induces the same act, and concave in $\mathbf{w}$, an affine function minus a maximum of affine functions. In this notation (4) is $\sum_k w_k \varphi(U_k(x,n)) - A(\mathbf{w}) - C(\mathbf{w}) - cn$.

2.3 Specifying the optimization problem in closed form

The objective (4) is defined for general payoffs and signals. For tractability I specialize to a quadratic payoff and Gaussian signals; which results depend on these forms is discussed in Section 3.7. Let the act be a scalar $a \in \mathbb{R}$ (for example, the scale at which a farmer adopts a new technology) with

$$u(a,\theta) \;=\; a(\theta-\chi) - \tfrac{\lambda}{2}a^2, \qquad \lambda > 0, \tag{8}$$

[22] The update takes effect when the weights are re-chosen, which is deliberation and begins the next stage; within a stage the weights stay at their step-(2) values, as in (4).

where χ is the marginal cost of the act and λ an adjustment cost; the example in Section 2.1 sets $\chi = 0$ and $\lambda = 1$. Subsequent signals are conditionally i.i.d. draws $y_i = \theta + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \nu_y^2)$, so mental model k's posterior precision after x and n signals is $P_k(n) = \tau_k + \nu_k^{-2} + n\nu_y^{-2}$, the variance of its step-(4) posterior mean viewed from step (2) is $\psi_k(n) := 1/P_k(0) - 1/P_k(n)$, increasing and concave in n, and the model-contingent value is

$$U_k(x,n) \;=\; \frac{(m_k - \chi)^2 + \psi_k(n)}{2\lambda}, \qquad \frac{\partial U_k}{\partial n} \;=\; \frac{1}{2\lambda\,\nu_y^2 P_k(n)^2} \;>\; 0. \tag{9}$$

In the baseline $\varphi(U) = U$, and (7) evaluates to the weighted variance of the model-implied posterior means, $A(\mathbf{w}) = \mathrm{Var}_{\mathbf{w}}(m)/2\lambda$ with $\bar{m} := \sum_k w_k m_k$. When a single alternative k is held, $A = w_k(1-w_k)D_k(x)$ with

$$D_k(x) \;:=\; \frac{(m_k - m_0)^2}{2\lambda}, \tag{10}$$

the *disagreement* between alternative k and the default: how far apart the two mental models' recommended acts are given the signal. It is not a parameter of the DM, since each posterior mean reads the initial signal through its own attention mapping. The DM's problem becomes

$$\max_{\mathbf{w},n} V(\mathbf{w},n) \;=\; \sum_{k=0}^{K} w_k U_k(x,n) \;-\; A(\mathbf{w}) \;-\; C(\mathbf{w}) \;-\; cn \quad \text{s.t.} \quad 0 \le w_k \le \overline{W}_k(x),\; k \ge 1,\;\; n \in \mathbb{Z}_{\ge 0}, \tag{11}$$

and, when a single alternative k is held at weight w_k,

$$V(w_k,n) \;=\; (1-w_k)\,U_0(x,n) + w_k\,U_k(x,n) \;-\; w_k(1-w_k)D_k(x) \;-\; \tfrac{\gamma}{2}w_k^2 \;-\; cn. \tag{12}$$

Weight on an alternative offers returns in the value gap $U_k - U_0$, at the hedging cost $w_k(1-w_k)D_k$ and the cognitive cost $(\gamma/2)w_k^2$; signals refine each mental model's recommendation through its within-model value of information ψ_k, at a price of c per signal; when the act responds to updated weights, they can also resolve the uncertainty between models. Since the hedging cost measures the model uncertainty the DM holds, problem (11) is the choice of how much model uncertainty to entertain, through the weights, and how much information to acquire, through the signals.

Discussion on beliefs and welfare. The weights are the DM's beliefs about which mental model applies, second-order beliefs in the sense of Klibanoff et al. (2005), and the objective is evaluated under them. They are chosen, akin to models of motivated beliefs, but only at or below their posteriors. Realized payoffs are determined by whichever model is true, so neglecting a correct alternative lowers them; what the DM buys with that loss is the cognitive cost of holding it. The hedging cost, the expected value of perfect information about the model, is what makes the objective convex in the weights, and it vanishes as she commits to one model.

The theory is positive. The DM optimizes given cognitive costs, so she is rational in that sense, but the costs make her choices depart from a Bayesian's, and the Bayesian posterior serves as the yardstick for measuring the departure, not as her objective. Welfare is measured by realized payoffs under whichever model is true, net of the cognitive costs, which are real. By that measure the loss from neglecting a correct alternative has a simple form: acting on the default when the alternative is right costs the disagreement D_k, the payoff gap between the two recommended acts, plus the value of the information she did not gather (Remark OA.1 in the Online Appendix). The policy statements later in the paper rest on this measure and on one feature of the constraint: an intervention that works through evidence can reduce neglect but cannot push weight above the posterior, so it moves the DM toward the Bayesian benchmark and never past it.

3 Main Results

This section presents the main results of the paper, which characterize the conditions under which the DM entertains alternative mental models and seeks the information required to explore them. I first characterize the thresholds in alternative activation, the curiosity failures that arise when the evidence constraint meets a complementarity between model weights and information acquisition, in which the DM fails to investigate interpretations that could change her decisions, and the initial signal required to clear the thresholds. I then show that curiosity can be defined endogenously as the value of resolving the model uncertainty that activation creates. Finally, I add a non-instrumental motive for curiosity through ambiguity aversion and endogenize the type of information acquired.

A property of the cognitive cost function simplifies the results in this paper.

Lemma 2 (Spreading attention weights across alternatives is costly)**.** *Fix total attention weight off the default, $W = \sum_{k\geq 1} w_k$. Then $C(\mathbf{w})$ in (3) is uniquely minimized by placing all of W on a single alternative, and the cost of any interior split strictly exceeds the concentrated cost by $(\eta - \gamma)\sum_{j<k} w_j w_k > 0$.*

The DM's problem thus has a built-in preference for entertaining one alternative at a time, the endogenous counterpart of a restriction imposed by assumption elsewhere in the model-switching literature (Ortoleva, 2012; Schwartzstein and Sunderam, 2021). When a single alternative k is held, the objective is (12): a comparison between alternative k and the default, with the disagreement $D_k(x)$, the value gap $\Delta U_k := U_k(x,n) - U_0(x,n)$, and the evidence cap $\overline{W}_k(x)$ all computed from the initial signal. The DM entertains the alternative for which the optimized value of (12) is highest, and the initial signal determines it through the same three objects. The next subsections characterize activation for a given alternative before returning to the interaction between alternatives in Section 4.2.

3.1 Alternative activation

Activation of alternative k is governed by a value condition and the evidence constraint. The *value condition* is that the value gap ΔU_k clear a threshold T_k, so that the alternative is worth enough to pay for holding both models.[23] The *evidence constraint* is that the weight on k cannot exceed $\overline{W}_k(x)$, its evidence cap, which is near zero when x is consistent with the default. Which form the value condition takes depends on the curvature of (12) in w_k, namely $2D_k - \gamma$, so T_k is stated regime by regime.

Proposition 1 (Activation thresholds)**.** *Fix n and an alternative k held on its own, and write $D_k = D_k(x)$ and $\overline{W}_k = \overline{W}_k(x)$.*

(i) (Thresholds.) *There is a threshold $T_k \leq D_k$ such that $w_k^* = 0$ if $\Delta U_k \leq T_k$ and $w_k^* > 0$ if $\Delta U_k > T_k$, and in either case $w_k^* \leq \overline{W}_k$.*

(ii) (Two regimes.) *The threshold and the form of the optimum depend on the curvature $2D_k - \gamma$.*

(a) (Graded regime, $\gamma > 2D_k$.) *$T_k = D_k$, V is strictly concave in w_k, and*

$$w_k^* \;=\; \text{med}\Big\{0,\; \frac{\Delta U_k - D_k}{\gamma - 2D_k},\; \overline{W}_k\Big\},$$

which is continuous and increasing in ΔU_k.

(b) (All-or-nothing regime, $\gamma < 2D_k$.) *V is strictly convex in w_k, so $w_k^* \in \{0, \overline{W}_k\}$, and $w_k^* = \overline{W}_k$ if and only if*

$$\Delta U_k \;>\; T_k \;=\; T_k(\overline{W}_k) \;:=\; D_k(1 - \overline{W}_k) + \tfrac{\gamma}{2}\overline{W}_k \;<\; D_k.$$

[23] "Worth" here means decision value, not optimism. Under (9), U_k rises with how decisive the action model k recommends is, $(m_k - \chi)^2$, and with what remains to be learned under it, $\psi_k(n)$. An alternative that reads the opportunity as harmful, and so calls for decisively opting out, can therefore carry a larger value gap than a lukewarm default.

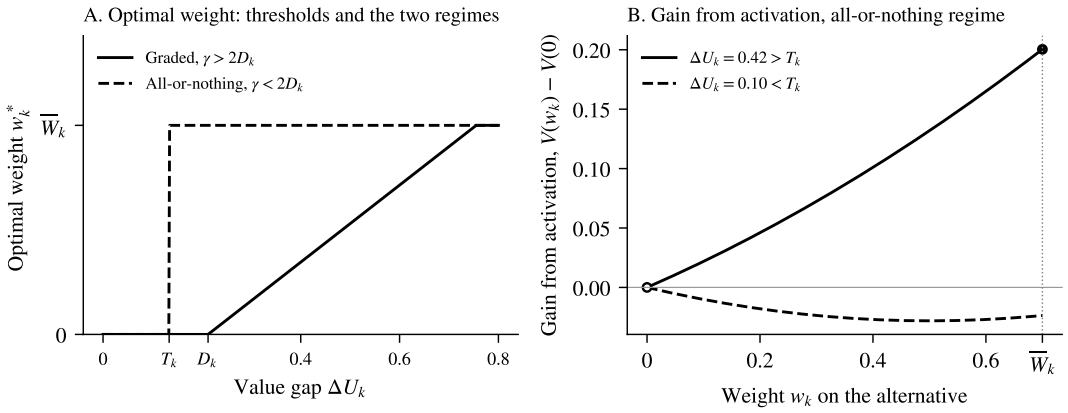


FIGURE 2. THRESHOLD EFFECTS IN ALTERNATIVE ACTIVATION

Notes: Both panels illustrate Proposition 1 with $D_k = 0.2125$ and evidence cap $\overline{W}_k = 0.70$. Panel A plots the optimal weight against the value gap: in the graded regime ($\gamma = 1.20 > 2D_k$) the threshold is $T_k = D_k = 0.2125$ and the weight rises continuously beyond it until it reaches the cap; in the all-or-nothing regime ($\gamma = 0.20 < 2D_k$) it is lower, $T_k(\overline{W}_k) = 0.134$, and the weight jumps from zero to the cap there. Panel B plots the gain from activation, $V(w_k) - V(0)$ from (12), in the all-or-nothing regime: the gain is convex in the weight, so for a value gap above $T_k(\overline{W}_k)$ (solid) the DM moves straight to the cap, and below it (dashed) any partial activation is worse than none.

> *The threshold is strictly below D_k: when the alternative and the default disagree enough, a decisive shift can pay even though no partial shift would. The threshold T_k is strictly decreasing in $\overline{W}_k$, and at the threshold w_k^* jumps discontinuously from 0 to $\overline{W}_k$.*

Both $\overline{W}_k(x)$ and $D_k(x)$ are functions of the initial signal, so the initial signal determines which alternatives can be activated and, for each, whether the weight on it moves continuously or in switches.

Figure 2 illustrates Proposition 1. In Panel A the threshold is the hedging cost of the first unit of weight: holding an alternative opens model uncertainty where there was none, at a marginal cost of D_k, so an alternative must be worth more than the disagreement it introduces, and one that disagrees sharply with the default is *harder* to activate partially. Panel B shows why the weight jumps in the all-or-nothing regime. The curvature $2D_k - \gamma$ compares two marginal costs of weight: the hedging cost falls as the weight rises, because a mixture is most uncertain when most evenly split, while the cognitive cost rises. When disagreement is large relative to γ, compromise is the costlier way to hold the alternative and the DM resolves her model uncertainty by switching; when it is small, attention weights are the costlier way and she adjusts at the margin. The evidence constraint bounds the weight in both regimes but works differently in each. In the graded regime it binds only once the value gap is large enough, and the threshold stays at D_k. In the all-or-nothing regime it binds at every activated optimum, since the switch is to the cap, and it shapes the threshold: a higher cap lowers $T_k(\overline{W}_k)$, because the hedging cost per unit of weight, $(1 - w_k)D_k$, falls as w_k rises toward 1, so a shift to a higher cap carries less model uncertainty per unit of weight. One consequence is that an isolated favorable observation that falls short of the threshold leaves no trace in the all-or-nothing regime, and only a small weight that does not grow in the graded one (Corollary OA.1 in the Online Appendix).

The two regimes are behavioral analogues of the gradual and abrupt transitions neuroscientists observe when an animal's environment is morphed step by step into another: in some experiments the hippocampal representation changes progressively (Leutgeb et al., 2005), in others it switches all at once to an independent code (Wills et al., 2005; Colgin et al., 2008), a switch that Sanders et al. (2020) interpret as Bayesian inference about which situation the animal is in. The analogy concerns the shape of the transition, not its neural implementation: in the brain, which pattern occurs depends on which inputs change and on prior experience, not on a comparison of payoffs.

3.2 Complementarity and curiosity failures

Proposition 1 holds the number of signals fixed. The DM chooses w_k and n together, and the two choices are complements: weight on an alternative raises the value of signals about it, and signals raise the value of weight. This subsection states the complementarity, derives the information acquisition it implies, and shows how, combined with the evidence constraint, it produces curiosity failures.

Proposition 2 (Complementarity)**.** *Under the maintained assumption $P_0(0) > P_k(0)$, the effect of an additional signal on the return to attention weight on alternative k, the cross-partial of* (4)*, is*

$$\frac{\partial^2 V}{\partial w_k \, \partial n} = \frac{\partial U_k}{\partial n} - \frac{\partial U_0}{\partial n} = \frac{1}{2\lambda \nu_y^2}\Big[\frac{1}{P_k(n)^2} - \frac{1}{P_0(n)^2}\Big] > 0,$$

since the alternative is the less precise model, $P_k(n) < P_0(n)$. Hence (w_k, n) are complements: w_k^* is nondecreasing in n and n^* is nondecreasing in w_k, and, treating n as continuous, the first-order condition in n is $\sum_{j=0}^{K} w_j \, \partial_n U_j = c$.

Weights and information are complements because the alternative has more left to learn: the marginal value of a signal falls with a model's precision. To see how the complementarity interacts with the evidence constraint, it helps to solve for the information the DM acquires in closed form. Unlike the example of Section 2.1, where inquiries refine both interpretations, I now consider signals that inform the alternative alone, which yields a simpler information threshold.

Assumption 1 (Alternative-specific signals)**.** The subsequent signals load on features the default mental model does not encode: under each alternative they follow $y_i = \theta + \varepsilon_i$, while under the default they are drawn from a density f_0 that does not depend on θ. The default's posterior over θ therefore does not update, $P_0(n) = P_0(0)$ and $\partial U_0 / \partial n = 0$, and f_0 is the default's likelihood in (6).

Under Assumption 1 the marginal benefit of a signal at $\mathbf{w} = 0$ is zero, so $n^* = 0$ under the default alone. Under Assumption 1, information seeking is thus absent under the default alone, however diffuse its prior, and grows with the attention weight the alternative receives: there is no point acquiring detail about a model that has not been activated.[24]

Corollary 1 (Closed-form information acquisition)**.** *Under Assumption 1, when alternative k is held on its own at weight w_k, the n-th signal is worth $w_k/(2\lambda \nu_y^2 P_k(n-1)P_k(n))$, which is decreasing in n. The optimal number of signals is therefore the largest n whose gain is positive, $P_k(n-1)P_k(n) < w_k/(2\lambda \nu_y^2 c)$, with ties broken against buying: zero for $w_k \le \underline{w}_k := 2\lambda \nu_y^2 c \, P_k(0)P_k(1)$, and thereafter nondecreasing in w_k and nonincreasing in c. Treating n as continuous gives the approximation $n^*(w_k) \approx \nu_y^2 \big[(w_k/2\lambda \nu_y^2 c)^{1/2} - P_k(0)\big]^+$. Every optimum (w_k^*, n^*) is a fixed point of two explicit best responses: the weight $w_k^*(n)$ of Proposition 1, which depends on n through the value gap $\Delta U_k(n)$, and the number of signals $n^*(w_k)$. An optimum exists, since the value of signals is bounded while their cost grows linearly in n. Because the two best responses are complements, there can be several fixed points, and the optimum is the one with the highest value.*

Curiosity failures. Much like collateral constraints in models of poverty traps (Galor and Zeira, 1993; Banerjee and Newman, 1993), the evidence constraint generates threshold effects in the exploration of alternative mental models. Corollary 1 contains an information threshold: no signal is bought until the weight on the alternative exceeds $\underline{w}_k := 2\lambda \nu_y^2 c \, P_k(0)P_k(1)$, a level that rises with the cost of signals and with the

[24] The assumption is used for two things only: the result that no information is acquired under the default alone, $n^* = 0$ at $\mathbf{w} = 0$, and the closed form for the number of signals in Corollary 1. The complementarity holds without it (Proposition 2), and Section 3.6 drops it and recovers the same structure.

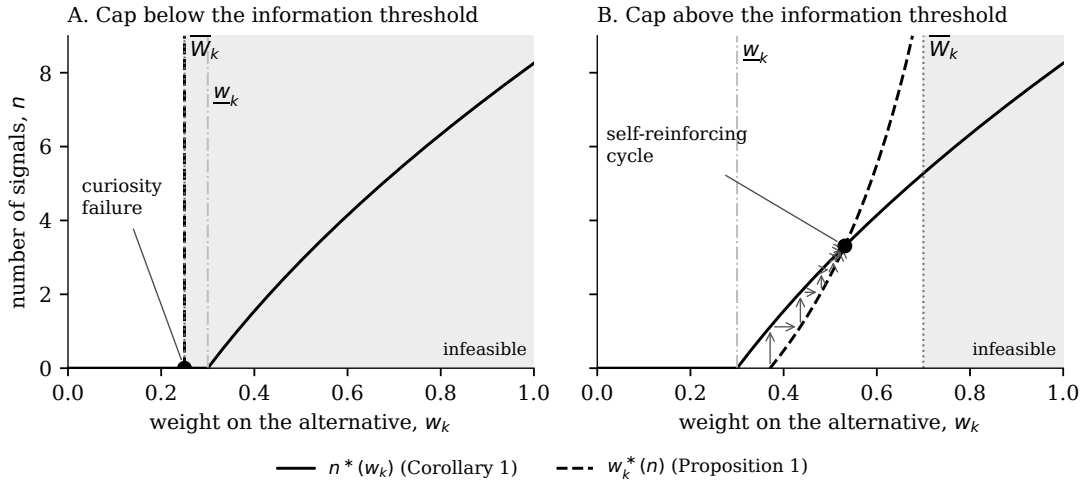


FIGURE 3. CURIOSITY FAILURES

Notes: Both panels plot, in the space of the weight on the alternative and the number of signals, the two best responses whose fixed point is the optimum of Corollary 1: the number of signals $n^*(w_k)$ given the weight (solid) and the weight $w_k^*(n)$ given the signals (dashed), in the graded regime of Proposition 1. Parameters are $\lambda = 1$, $v_y^2 = 10$, $P_k(0) = 1$, $c = 0.015$, $D_k = 0.2125$, $\gamma = 1.20$, and $\Delta U_k(0) = 0.50$; n is treated as continuous, for which the information threshold is $\underline{w}_k = 2\lambda v_y^2 c P_k(0)^2 = 0.30$. The shaded area is infeasible under the evidence constraint. Panel A sets the cap at $\overline{W}_k = 0.25 < \underline{w}_k$: the weight is clamped at the cap, where no signal pays, and the fixed point is $(0.25, 0)$. Panel B sets $\overline{W}_k = 0.70 > \underline{w}_k$: the arrows trace the complementarity of Proposition 2, converging to $(0.53, 3.3)$. In both panels the point shown is the global optimum.

alternative's own precision.[25] By Proposition 1, the weight cannot exceed the evidence cap $\overline{W}_k(x)$, and within a stage the initial signal fixes the cap. If $\overline{W}_k(x) \leq \underline{w}_k$, therefore, the alternative may be activated but is never investigated: $n^* = 0$, its posterior stays imprecise, and by (6) its cap never rises. This is a *curiosity failure*: the DM fails to investigate an interpretation that could change her decision, even when the relevant information is available, however promising the alternative and even if it is right. The friction is endogenous to her interpretation. In a model of costly information or experimentation, the DM recognizes which evidence would be valuable and declines it only when it costs more than it is worth; here she can afford it but does not take the alternative seriously enough to recognize its value, because the alternative that would make it valuable carries too little weight. Her ignorance of the alternative prevents the investigation that would end it. If instead $\overline{W}_k(x) > \underline{w}_k$ and the alternative is activated at a weight above $\underline{w}_k$, the complementarity of Proposition 2 starts a self-reinforcing cycle: the weight justifies signals, the signals raise the cap through (6) when the alternative fits them, and at the next stage the higher cap admits more weight. Small differences in the initial signal can thus produce sharply different outcomes. Figure 3 illustrates with the two best responses of Corollary 1 for the same alternative under two evidence caps: below the information threshold the DM activates the alternative at the cap but buys no signal (Panel A), and above it the best responses meet at an interior fixed point with $n^* > 0$ (Panel B).

As an illustration, consider parents deciding whether to engage with a school-based agricultural education program. In a field experiment in Liberia (Lee, 2026), parents offered no further information largely avoided the program, while parents shown videos and field days presenting its distinguishing features visited school farms and met teachers on their own, and their households adopted more improved techniques. Through the lens of the theory, parents in both arms start from the same default, a typical school garden that uses students as farm labor. A program that fails to distinguish itself from it leaves the cap of the alternative reading below the level at which investigating it pays, so the alternative is never investigated and its cap never rises. The

[25] Treating n as continuous yields the same curiosity failures, with threshold $2\lambda v_y^2 c P_k(0)^2$, where the approximation in Corollary 1 turns positive; discreteness only raises it to the cost of a whole first signal.

videos and field days change the initial signal, lifting the cap above that level, after which the complementarity takes over and parents seek the evidence themselves.

3.3 The impact of the initial signal

Given the central role of the initial signal, what type of initial signals can clear the threshold? Not simply those that supply more observable features that fit the alternative: the evidence constraint rests on relative likelihoods, and a signal is a vector of features that each mental model reads through its own attention mapping, so how evidence is framed is central. Because (1) is linear Gaussian, each Bayes factor decomposes into three interpretable channels, which turns the evidence constraint into a design instrument. Under (1), $x \sim \mathcal{N}(\mu_k b_k,\ \Sigma + \sigma_k^2 b_k b_k')$ under mental model k, and $L_k(x)$ is that density evaluated at x (Lemma OA.3 in the Online Appendix).

Proposition 3 (Three-channel decomposition of the evidence constraint)**.** *Define, for each mental model* $k = 0, 1, \ldots, K$*,*

$$\underbrace{R_k := x'\Sigma^{-1}x - s_k^2/\nu_k^2}_{\textit{off-attention residual}}, \qquad \underbrace{q_k := \frac{(s_k-\mu_k)^2}{\sigma_k^2+\nu_k^2}}_{\textit{standardized surprise}}, \qquad \underbrace{c_k := \log\left(1+\frac{\sigma_k^2}{\nu_k^2}\right)}_{\textit{Occam factor}}.$$

Then $R_k \geq 0$ *and, for each alternative k,*

$$\log \Lambda_k(x) \;=\; \tfrac{1}{2}\Big[\underbrace{(R_0 - R_k)}_{\textit{encoding}} + \underbrace{(q_0 - q_k)}_{\textit{surprise}} + \underbrace{(c_0 - c_k)}_{\textit{Occam}}\Big]. \tag{13}$$

Consequently:

- *(i) The encoding channel vanishes identically if and only if* $b_0 \propto b_k$*.*
- *(ii) The surprise channel is the only channel that depends on the prior means: it is positive if and only if the initial signal, read through each model's attention mapping, lies further from the default's prior mean than from the alternative's in standardized units,* $q_0 > q_k$*.*
- *(iii) The Occam channel is strictly negative whenever* $\sigma_k^2/\nu_k^2 > \sigma_0^2/\nu_0^2$*, in particular whenever the alternative has the default's loading,* $b_k = b_0$*, and the more diffuse prior.*

$\overline{W}_k(x)$ *is increasing in each of the three channel differences of alternative k, decreasing in each channel difference of every other alternative, and equals the Bayesian posterior model probability under a uniform prior over mental models.*

The decomposition identifies three margins on which a signal can raise the evidence cap of an alternative. The first is *encoding*, the channel $R_0 - R_k$: the signal carries features that the alternative reads but the default ignores. The second is *surprise*, the channel $q_0 - q_k$: the alternative is less surprised by the part of the signal it reads than the default is by the part it reads, as when it better predicts an observation that is extreme on a feature both mental models read. The third is *specificity*, the Occam channel $c_0 - c_k$. A marginal likelihood penalizes a model for spreading predictive mass, and c_k is the Occam factor of Bayesian model comparison (MacKay, 1992; Jefferys and Berger, 1992; Kass and Raftery, 1995; Ba, 2026). Every alternative therefore begins at a handicap relative to the more precise default, and the handicap shrinks as the alternative's prior approaches the default's in precision. Since c_k is increasing in σ_k^2, an optimistic but vague alternative ("the program might really help") is harder to justify than a specific one ("students' scores in agriculture rise by about eight points"), even when the vague version has the higher mean. Precision has two limits. A narrower alternative whose predictions miss loses more likelihood than a vague one. And precision lowers value: a

more precise prior leaves less to learn, so $\psi_k(n)$ and with it the value gap ΔU_k fall. Narrowing an alternative therefore raises its evidence cap but can weaken the value condition, and it helps most when the cap binds.

The three-channel decomposition suggests an implication for designing information interventions via the initial signal, and a limit. The implication is that exploration can be stimulated while holding constant the observable features the default reads, by presenting features that only an alternative reads (encoding) or by stating the alternative as a specific, checkable claim consistent with what is observed (specificity). An intervention can thus activate an alternative without adding to the evidence the default already reads. The intervention of Hanna et al. (2014) can be read in these terms. Seaweed farmers with years of experience had not learned that pod size mattered for yield, although the relationship was present in their own plots. A summary of their own trial data, organized by pod size, then changed their practices although it contained no observation they had not already made; in the present framework, the summary is a signal that an alternative reads and the default does not. The limit is that the initial signal works only through the mental models on the menu: an intervention can target and activate an alternative the DM can entertain, but a feature no mental model reads leaves every channel, and hence every evidence cap, unchanged.

3.4 A formal definition of curiosity

Curiosity—the drive to seek the information required to explore new perspectives—requires no additional modeling: it emerges from the setup. Once an alternative carries weight, the DM must choose one act for two mental models that recommend different ones, and the value she loses to that compromise, the hedging cost, is by construction what she would gain from learning which model is right. A drive to resolve model uncertainty via information acquisition is therefore already present in the objective. This drive shares its structure with the information-gap definition of curiosity in psychology, a gap between what one knows and what one wants to know, which exists only once a question has been posed (Loewenstein, 1994). While Loewenstein (1994) gives the information gap a perceptual explanation—a cognitively induced deprivation that arises from perceiving the gap—this paper models the motive to close it as instrumental: closing the gap is valuable because it changes what the DM does. Two reasons support this choice. First, many non-instrumental motives can be read instrumentally: a rational analysis of curiosity treats a person as exploring what most improves the usefulness of her knowledge (Dubey and Griffiths, 2020), and curiosity responds experimentally to the stakes and future usefulness of what is to be learned (Kobayashi et al., 2019; Liquin and Lombrozo, 2020). Second, the instrumental motive is the one an expected-utility maximizer already has (Blackwell, 1953), so a baseline built on it depends on no individual differences in tastes, and a non-instrumental motive can be added on top of it (Section 3.5). Here the question is posed by activation, the gap is the disagreement between the mental models held, and its pull is the hedging cost, which accordingly serves as the definition.

Definition 1 (Curiosity)**.** At weights $\mathbf{w}$, *curiosity* is the value of resolving which mental model is right, which is the hedging cost (7),

$$A(\mathbf{w}) \;=\; \frac{\mathrm{Var}_{\mathbf{w}}(m)}{2\lambda},$$

the component of the value of information that comes from discriminating between the models rather than from refining either.

Curiosity is thus a function of the weights, determined jointly with the information the DM acquires, and it shows in behavior through that information: how much and along which features. Three properties follow. Curiosity is zero under the default alone, however diffuse that model is. When a single alternative k is held it equals $w_k(1-w_k)D_k$, inverted-U in the weight and greatest when the DM is least sure which model applies, the intensity profile of the information-gap account, in which curiosity peaks at intermediate confidence; Kang et al. (2009) find this shape for trivia questions, where no decision is at stake, and here it arises for questions

that bear on one. And it values only signals that discriminate between the mental models, not signals that refine both, so it is directed at the observations that would settle which model is right. The last property separates curiosity from the value of information as such.[26]

The next result completes the case for Definition 1. The objective (4) holds the weights fixed within a stage, the natural reading when the DM must act before she can deliberate again, since re-weighting is itself deliberation; suppose instead that the act is chosen at the weights updated by (6). Let $\bar{m}_n := \sum_{k=0}^{K} w_{k,n} m_{k,n}$ be the mixture posterior mean after the signals, and let $A_n := \mathrm{Var}_{\mathbf{w}_n}(m_{\cdot,n})/(2\lambda)$ be the hedging cost the DM faces at step (4), the analogue of $\mathrm{Var}_{\mathbf{w}}(m)/2\lambda$ at the updated weights and means.

Proposition 4 (The value of information with updating)**.** *Under* (8) *with φ linear and the update* (6)*, the ex-ante value to the DM of the n signals, evaluated under the mixture* $\mathbf{w}$*, is*

$$\mathrm{VOI}(\mathbf{w},n) \;=\; \frac{\mathrm{Var}_{\mathbf{w}}(\bar{m}_n)}{2\lambda} \;=\; \underbrace{\frac{\sum_{k=0}^{K} w_k\,\psi_k(n)}{2\lambda}}_{\text{within-model refinement}} \;+\; \underbrace{A(\mathbf{w}) - \mathbb{E}_{\mathbf{w}}[A_n]}_{\text{resolution of model uncertainty}} \qquad . \tag{14}$$

(a) The resolution component vanishes whenever the DM holds a single mental model, $w_k = 1$ for some k, and converges to the hedging cost, $A(\mathbf{w}) - \mathbb{E}_{\mathbf{w}}[A_n] \to A(\mathbf{w})$ as $n \to \infty$: in the limit, information recovers exactly the cost of holding several mental models at once.

(b) Let a single alternative k be held against the default, with $w_k \in (0,1)$, and let θ_0 denote the parameter generating the subsequent signals. The following limits hold almost surely as $n \to \infty$.

(i) (Shared signals.) *If both mental models read the signals as informative about θ, then $m_{k,n} - m_{0,n} \to 0$ and $w_{k,n} \to w_k\Lambda_k^\infty/(1 - w_k + w_k\Lambda_k^\infty) \in (0,1)$, where $\Lambda_k^\infty := g_k(\theta_0)/g_0(\theta_0)$ is the ratio of the two models' posterior densities of θ at θ_0 after the initial signal.*

(ii) (Alternative-specific signals.) *Under Assumption 1, if the signals are drawn from $y_i = \theta_0 + \varepsilon_i$, the process the alternative specifies, then $w_{k,n} \to 1$; if they are drawn from the density f_0 that the default assigns them, and f_0 is not a Gaussian with variance ν_y^2, then $w_{k,n} \to 0$.*

Part (a) makes precise what the introduction calls resolving the model uncertainty that activation creates: the hedging cost is by construction the value of learning which mental model is right, and enough signals deliver that value in full. Part (b) separates two things "resolving model uncertainty" can mean. With shared signals the models come to recommend the same action, so the hedging cost disappears, but the weight settles strictly inside the unit interval: the question of which model is right is never settled. With alternative-specific signals one model wins, the weight goes to 0 or 1, and the DM keeps one mental model.

Comparing (14) with (11), choosing the act at the updated weights adds the resolution component to the baseline, which credits signals with refinement only. The information threshold survives: under Assumption 1, with one alternative held, the resolution component is at most $w_k(1-w_k)D_k$, so no signal is bought at a weight below $c/[D_k + 1/(2\lambda P_k(0))]$. A curiosity failure therefore arises under either criterion when the evidence cap falls below the corresponding threshold.

3.5 Ambiguity aversion

Section 3.4 set aside the non-instrumental motive for curiosity: the cognitively induced deprivation that Loewenstein (1994) attributes to perceiving an information gap, which makes resolving a question valuable even when the answer would not change what one does. A natural source of this motive is ambiguity aversion. A strictly concave φ, ambiguity aversion in the smooth ambiguity model (Klibanoff et al., 2005),

[26] Given the weights, the value of information here is an expected-utility object: the value of refining the mixture belief plus the expected value of perfect information about which model is right. What is not an expected-utility object is activation, the choice of the weights: a Bayesian who holds two models is always curious; the DM here is curious only once she has chosen to hold a second one.

makes holding two mental models costly in itself. This subsection shows that it plays two roles. It makes alternatives harder to activate, exactly for any strictly concave φ; and once the DM also cares about when model uncertainty is resolved, it makes settling which mental model is right valuable for its own sake.

Proposition 5 (Ambiguity aversion and activation)**.** *Let φ be strictly concave, normalized so that $\varphi'(U_0) = 1$, which puts values and costs in the same units at the default, and let $n = 0$.*

(i) (Exactly, for any φ.) *For every w_k, the gain from holding alternative k at weight w_k rather than the default alone is at most its value under linear φ, and it falls further with ambiguity aversion in the sense of Pratt (1964). The first unit of weight is worth $\varphi(U_k - D_k) - \varphi(U_0) \leq \Delta U_k - D_k$, which has the sign of $\Delta U_k - D_k$. Hence ambiguity aversion never activates an alternative that linear φ leaves inactive and never blocks one with $\Delta U_k > D_k$: it can block activation only in the all-or-nothing band $T_k(\overline{W}_k) < \Delta U_k \leq D_k$, and a more ambiguity-averse DM activates a weakly smaller set of alternatives.*

(ii) (Locally.) *Let $\xi := -\varphi''(U_0)$, the coefficient of ambiguity aversion at the default under the normalization. To second order in the difference between the models' predictions, the gain from holding k at weight w_k is that of* (12) *with the value gap ΔU_k replaced by $\Delta \tilde{U}_k := \Delta U_k - \frac{\xi}{2}(\Delta U_k)^2$: ambiguity aversion acts as a reduction in the value gap that grows with the stakes, and it leaves the curvature of the objective in w_k, and so the regime, unchanged at this order.*

Ambiguity aversion thus makes alternatives harder to activate, but only in the all-or-nothing band: an alternative whose value gap exceeds the disagreement is activated whatever the DM's attitude toward model uncertainty, though locally at a lower weight, and the force of ambiguity aversion grows with the stakes $(\Delta U_k)^2$.

On its own, ambiguity aversion does not make resolving which mental model is right valuable for its own sake. If the DM evaluates information model by model, as the recursive smooth ambiguity criterion of Klibanoff et al. (2009) does, a signal that reveals the model is worth only what it saves in hedging, which is zero when the models recommend the same act. A non-instrumental motive appears once she also has an attitude toward when model uncertainty is resolved, as recursive models of ambiguity aversion generally imply (Strzalecki, 2013; Li, 2020). Suppose she values a signal by the certainty equivalent she expects to hold once it arrives, treating its realization as risk under her current beliefs: the analogue, for model uncertainty, of a preference for early resolution (Kreps and Porteus, 1978). As in Proposition 4, the act is chosen at the updated weights. Write $G := \varphi^{-1}(\max_a \sum_k w_k \varphi(\mathbb{E}_k u(a,\theta)))$ for the certainty equivalent of the single act, $\mathrm{CE} := \varphi^{-1}(\sum_k w_k \varphi(U_k))$ for that of knowing the model, $\pi := \bar{U} - \mathrm{CE}$ for the *ambiguity premium*, with $\bar{U} := \sum_k w_k U_k$, and G_n, CE_n, π_n, $\bar{U}_n$ for their values after the signals.

Proposition 6 (Ambiguity aversion and the value of resolution)**.** *Let φ be strictly concave and let a signal be worth $\mathbb{E}_{\mathbf{w}}[G_n] - G$, the expectation taken under the mixture predictive.*

(i) The value of any signal decomposes as

$$\mathbb{E}_{\mathbf{w}}[G_n] - G \;=\; \underbrace{\mathbb{E}_{\mathbf{w}}[\bar{U}_n] - \bar{U}}_{\textit{refinement}} + \underbrace{H - \mathbb{E}_{\mathbf{w}}[H_n]}_{\textit{resolution}} + \underbrace{\pi - \mathbb{E}_{\mathbf{w}}[\pi_n]}_{\textit{non-instrumental}},$$

where $H := \mathrm{CE} - G \geq 0$ is the hedging cost in certainty-equivalent units. Under linear φ the last term vanishes and the decomposition is (14)*.*

(ii) With alternative k held at weight w_k, a signal that reveals which mental model is right, without changing either model's predictions, is worth $H(w_k) + \pi(w_k)$. When the models recommend the same act, $H = 0$ and its whole value is the premium $\pi(w_k)$, which is strictly concave in w_k, positive except at $w_k \in \{0,1\}$, increasing in $|\Delta U_k|$ and in ambiguity aversion in the sense of Pratt (1964), and to second order equal to $\frac{\xi}{2} w_k(1-w_k)(\Delta U_k)^2$, with ξ as in Proposition 5.

Part (ii) is the non-instrumental motive of the information-gap account (Loewenstein, 1994; Golman and Loewenstein, 2018), confined here to questions that bear on payoffs: the DM values learning which mental model is right even when the answer would not change what she does, most when she is least sure which model applies, and in proportion to the stakes. The two attitudes play distinct roles. Ambiguity aversion alone decides whether a question is opened; the attitude toward its resolution decides whether, once open, it is pursued for its own sake. The $w_k(1-w_k)$ shape shared by the hedging cost and the premium is the one that the source function of Abdellaoui et al. (2011) and Baillon et al. (2018) attaches to perceived ambiguity, so interior weights, rather than a diffuse single prior, are the object those indices pick up.

Because the premium grows with $(\Delta U_k)^2$, both effects are largest for questions with high stakes: such questions are harder to open, but once open they are pursued more actively. Where stakes are high, interventions that activate an alternative therefore become more effective relative to those that lower the cost of information. For example, when farmers suspect that fertilizer bought on credit may be fake (Bold et al., 2017; Michelson et al., 2021), the model predicts that side-by-side plots grown from the same bag or nutrient tests at the point of sale do more than credit or subsidies that only make a trial cheaper, and that small trial packs, by lowering the stakes, make the question easier to open. The local comparison across stakes requires that absolute ambiguity aversion ξ not fall in proportion to them; the exact results need no such condition, since any increase in ambiguity aversion weakly shrinks the set of alternatives activated and raises the value of settling the question.

3.6 Endogenous exposure: where the decision maker seeks information

So far the DM has chosen only how many signals to acquire, and under Assumption 1, which facilitates closed-form expressions, those signals bear on the alternative alone. Curiosity also manifests in the type of information the DM seeks once her exposure to signals is made endogenous. This subsection relaxes Assumption 1 and allows signals to bear on the default as well as on the alternative. From here on, each signal also has a direction d, $y = d'\tilde{x} + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \nu_y^2)$, so that the DM chooses what information to look at in addition to how much to acquire. Directions may include those the default reads. She chooses the direction first and then how many observations to make. A school visit that looks at crop yields is one d; one that looks at what students can do is another. With a single alternative k held at weight $w = w_k$, and observations valued as in Proposition 4, with the act chosen at the updated weight, write $\delta := b_k m_k - b_0 m_0$ for the difference in the two models' predictions, $\Omega_0 := \Sigma + b_0 b_0'/P_0(0)$ for the default's predictive covariance, and

$$v(w) \;:=\; \underbrace{(1-w)\,\frac{b_0}{P_0(0)} + w\,\frac{b_k}{P_k(0)}}_{\text{refinement}} \;+\; \underbrace{w(1-w)\,(m_k - m_0)\,\delta}_{\text{resolution}}.$$

Proposition 7 (Optimal direction of exposure)**.** *To leading order as the observation becomes weak relative to the DM's prior uncertainty, $\nu_y^2 \to \infty$, the value of an observation along d is proportional to $(d'v(w))^2/(d'\Omega_0 d + \nu_y^2)$. Among directions normalized by $d'\Omega_0 d = 1$, the DM therefore looks along*

$$d^o(w) \;\propto\; \Omega_0^{-1}\, v(w).$$

(i) At $w = 0$ she looks through the default's lens, $\Omega_0^{-1} b_0$, and at $w = 1$ through the alternative's, $\Omega_0^{-1} b_k$.

(ii) In between, the pull toward the direction of disagreement $d^* \propto \Omega_0^{-1}\delta$, the direction that most reduces the DM's model uncertainty, $w(1-w)$, is proportional to $w(1-w)(m_k - m_0)$: it is strongest when model uncertainty is greatest and when the answer would change what she does, and it vanishes when the two models recommend the same action, $m_k = m_0$, however much their predictions differ.

The proposition turns the choice of what to look at into a tradeoff between learning more within the

mental models she holds (refinement) and learning which of them is right (resolution), and the weight on the alternative sets the terms. When the alternative carries little weight, she looks where the default looks, and little, because the precise default has little left to learn and there is little model uncertainty to resolve: seeing what one expects to see is the optimal use of attention for someone with little model uncertainty. The value of an observation rises with the weight on the vaguer alternative, through both terms, which is the complementarity of weights and information in another form. As model uncertainty grows, she is pulled toward the features on which the mental models disagree, and once the alternative dominates, she looks through its lens to learn how to act on it. The resolution term is curiosity at work. It vanishes under the default alone and when both mental models recommend the same action, so curiosity targets the disagreements that matter for what she does, and it mirrors the directedness of curiosity in the information-gap account (Loewenstein, 1994): the gap is the predictive disagreement between the mental models held. Information gathered through the default's lens at low weight is routine refinement, not curiosity; what is absent under the default alone is information about which mental model is right.

The example of Section 2.1 treated every inquiry as a noisy report of the same return. Once the farmer can choose which evidence to seek, her interpretation also determines the direction of her search. Return to the farmer who sees a neighbor's good harvest. Her default attributes it to the weather; an alternative attributes it to the new seeds he planted. If she credits the weather, she watches the weather: rainfall data refine the model she holds and are absorbed by it, and nothing she sees there bears on whether the seeds matter. Once she suspects the seeds, she tries them on a plot of her own, since how that plot responds is the observation on which the two interpretations disagree and the one that would change what she does. Once she has adopted them, she stops testing and studies how to grow them well on her own farm, refining the alternative as she once refined the default. Likewise, for the school-based agricultural education program in Section 3.2: a parent who takes the school farm for a typical school garden only observes that students are farming; one who starts to view it as an educational program observes students' management capabilities; and one who views it as an integrative program of both agricultural extension and education assigns plots for students to try new farming practices.

Ambiguity aversion and the direction of search. Under the criterion of Proposition 6, the non-instrumental component also changes where the DM looks. A signal that discriminates between the models without refining either moves the weight as a martingale, and since the premium is concave in the weight, its non-instrumental value is nonnegative for every concave φ. A signal that refines both models without discriminating between them spreads the value gap and, locally, has negative non-instrumental value. Ambiguity aversion therefore tilts search toward ending model uncertainty rather than refining the mental model the DM favors (Proposition OA.4 in the Online Appendix).

Revealing mental models from choices. Read in reverse, the results of this subsection offer a way to observe what is otherwise hidden. Because the mental models a person holds direct her search, the signals she seeks and the amount she will pay for them help reveal those models. The direction of search helps reveal them qualitatively, as above. Valuations do so quantitatively, given model-contingent beliefs. Under the baseline criterion, the value of signals that only the alternative reads is $w_k\,\psi_k(n)/2\lambda$ (Corollary 1), linear in the weight, so, given each model's precision, their valuation bounds w_k. When the act responds to updated weights (Proposition 4), a signal that settles which model is right is worth the hedging cost $A(\mathbf{w})$, so its valuation measures curiosity directly. Valuations also detect a curiosity failure that search alone cannot: below the information threshold no signal is bought, but the valuation of an alternative-specific signal remains positive.

Remark (revealed mental models in practice). Digital farming advisory services can offer farmers messages that only one mental model reads, such as a rainfall record or a report on the new seeds, at randomized prices. The aim is not to recover a farmer's mental models but to detect how an intervention shifts them:

changes in her demand for the same message before and after the intervention track changes in her weight on the mental model that reads it. AI assistants add a second record, since shifts in what farmers ask after an intervention, for example from the weather to the seeds, reveal shifts in which features their mental models read (Proposition 7). In both settings, these shifts should predict whether a farmer then tries the new seeds.

3.7 Robustness

Four aspects of the analysis deserve scrutiny: the quadratic-Gaussian specialization, the DM's myopia across stages, the relation to models of experimentation, and whether the framework generates unique observable implications relative to existing Bayesian and behavioral benchmarks.

Functional forms. The evidence constraint and the crowding among alternatives in (5) are Bayes' rule and use no functional form. The two regimes need only that one act be chosen before the model is known: with a linear φ, the value of that act is a maximum of functions affine in the weight, hence convex, so the objective is a difference of convex functions, graded where γ exceeds the curvature of the value and all-or-nothing where it falls short; the quadratic payoff makes that curvature the constant $2D_k$ (Remark OA.2 in the Online Appendix). Curiosity failures need only a bounded payoff: the value of n signals at weight w_k is at most $w_k B_k$, where B_k is the most the signals could ever add to the value of acting on the alternative, so no signal is bought below the weight c/B_k, and an alternative whose cap lies below it is never investigated whatever its value (Proposition OA.2 in the Online Appendix). What the quadratic-Gaussian forms add is exactness: thresholds, the number of signals, and the complementarity of Proposition 2 as two-sided statements in closed form.

Myopia. Nor do curiosity failures rest on myopia. In a two-stage forward-looking extension, the DM would value a signal also for the weight it licenses tomorrow, but under her beliefs the cap cannot be expected to rise: it drifts up if the alternative is right and down if the default is. In the graded regime the value of a higher cap is concave, so this gamble is worth nothing in expectation and only the refinement value of reusing the signals remains, which lowers the information threshold by at most a factor of two; in the all-or-nothing regime the gamble has value bounded by how far the alternative's odds fall short of those needed to activate it (Proposition OA.3 in the Online Appendix). Curiosity failures therefore survive this extension: the extra incentive can tip only alternatives already close to activation.

Relation to bandit models. A bandit model can also generate exploration behavior, and with suitable modifications it can produce failures to explore as well. The present framework nonetheless differs from bandit models in two ways. First, a bandit asks whether to experiment with an uncertain option; here, the prior question is whether the DM takes an alternative interpretation seriously enough to recognize that investigating it could pay. Second, in a bandit the decision maker declines to experiment when the expected future benefit of learning falls short of its immediate cost. Here, an alternative can be promising, yet the evidence constraint limits the weight she can place on it, and because weight and information are complements, that low weight makes investigation unattractive. The DM thus has the capacity to explore that a bandit agent has, but it is subject to a cognitive limit that itself moves with what she observes.

Observable implications. Curiosity failures do not by themselves distinguish the framework from Bayesian model averaging with costly signals. A decision maker with the same mental models, signals, and payoffs, whose weights equal the posteriors rather than being chosen, also acquires no information about an alternative whose posterior lies below the threshold of Corollary 1. The distinguishing implications concern the impact of a plausible mental model that the DM would not act on and whose recommended act differs from the default's, so that any weight on it would show in what she does. Consider varying two of its features independently: its plausibility, measured by its likelihood L_j, and its recommended act. A mental model that promises no more

than the default, $\Delta U_j \leq 0$, is never activated, whatever its plausibility (Proposition 1). Its recommended act therefore has no effect on the DM's act, and no signal about it has value. Its likelihood nonetheless enters the denominator of every other cap in (5): a more plausible model of this kind lowers the cap on a promising alternative. Whether search falls depends further on that alternative: if it is active and its cap binds, the lower cap lowers its weight and, by Proposition 2, can lower the information acquired about it; if it is inactive or its cap does not bind, search is unchanged.

TABLE 1. THE IMPACT OF A PLAUSIBLE MENTAL MODEL THE DM WOULD NOT ACT ON

	This paper	Bayesian model averaging	Limited attention	Motivated beliefs
A more plausible model can reduce search for a promising alternative	Yes	Yes	No	No
Its recommended act moves the DM's act	No	Yes	No	Possibly
The DM seeks information about it	No	Possibly	No	No

Notes: The model recommends an act different from the default's. Entries describe the benchmark mechanisms as specified in the text, not every extension of them. In the first row, search falls only when the promising alternative is active and its cap binds.

Table 1 compares these predictions with three benchmarks. Under Bayesian model averaging, the same fragmentation arises, but only because the decision maker holds this model at its posterior; her act therefore moves toward its recommendation, and she may acquire information about it. Under limited attention (e.g., Schwartzstein, 2014; Matějka and McKay, 2015), a model that receives no attention weights has no effect. Under motivated beliefs without an evidence cap (e.g., Brunnermeier and Parker, 2005), an unappealing model may be discounted regardless of its plausibility. Among these benchmarks, only the present framework yields the joint pattern in which the model's plausibility affects search while its recommendation does not affect action. This joint pattern does not depend on the quadratic-Gaussian specification. The joint pattern is also testable under minimal conditions: the model's recommended act must differ from the default's, as assumed above, so that hedging under Bayesian model averaging is observable, and the variation in its plausibility must leave the evidence on the promising alternative unchanged. In the Gaussian specification such variation is feasible. A feature of the initial signal that only this model reads leaves every other alternative's Bayes factor unchanged (Proposition 3). That feature and the model's prior mean each move both its likelihood and its recommended act, in proportions that differ whenever the prior mean is nonzero, so adjusting the two jointly varies either one locally while holding the other fixed (Lemma OA.4 in the Online Appendix). Section 4.2 further elaborates on how this fragmentation among alternatives affects curiosity and pitches the framework against behavioral models of associative memory.

4 Behavioral Effects

Standard decision problems take the DM's representation of the situation as given. Associative-memory models, building on the psychology of memory (Howard and Kahana, 2002; Kahana, 2012), instead make that representation an outcome of recall: what comes to mind depends on its similarity to the situation, the cues present, and competition from other memories, and beliefs and choices follow from what is retrieved. One group of studies shows how recall shapes beliefs: people judge probabilities by the scenarios that come to mind (Gennaioli and Shleifer, 2010; Bordalo et al., 2023), imagine the future by drawing on memory (Bordalo et al., 2025), and react to news through cue-driven recall (Enke et al., 2024). Another shows how recall shapes choice: recalled experiences set the norm against which an option is judged (Bordalo et al., 2020), and a problem is recognized by retrieving the most similar category of past problems (Bordalo et al., 2026). Framing and context thus matter because they change what comes to mind, and so which problem the DM believes she is solving.

The framework here shares that premise. Associative memory supplies the default and the menu, and the mental model in use determines which features the DM attends to and how she reads them. What, then, do we gain from modeling the DM's capacity to be curious, on top of a default classification based on associative memory? Because she can seek information and reason beyond what associative memory supplies, the model helps characterize conditions under which two hallmark behavioral effects of associative-memory models, context effects and interference effects, persist. Where evidence cannot be sought, both persist as those models describe.

4.1 Context effects

In psychology, a context effect is the finding that cues present when an experience is stored and when it is recalled shape what is remembered, even when the cues are irrelevant to the task (Tulving and Thomson, 1973; Godden and Baddeley, 1975; Smith and Vela, 2001); the temporal context model formalizes how context drives retrieval (Howard and Kahana, 2002; Kahana, 2012). Economic associative-memory models build on this finding: a payoff-irrelevant cue changes which experiences or representations come to mind, and so changes beliefs and choices (Gennaioli and Shleifer, 2010; Bordalo et al., 2020; Enke et al., 2024; Bordalo et al., 2026). For example, a consumer sees a jar of jam in plain packaging as a routine purchase and focuses on its price, but the same jar in fancy packaging becomes a treat and draws her attention to its quality.

Here a contextual cue, such as where, when, or by whom a situation is presented, is a component of the initial signal x like any other: it is evidence only under a mental model that reads it. Since the default is taken as given (Section 2), this subsection concerns cues that leave the retrieved default unchanged; a cue that changes it acts through retrieval, which I leave to future research. The encoding channel of Proposition 3 determines when context matters. Call a component ζ of the initial signal *payoff-irrelevant under the default* if the default does not read it, $b_0'\Sigma^{-1}\zeta = 0$.

Corollary 2 (Context effects)**.** *Let x and $x+\zeta$ be two observations, where ζ is payoff-irrelevant under the default.*

(i) If no mental model on the menu reads ζ, that is, $b_k'\Sigma^{-1}\zeta = 0$ for every k, then $\Lambda_k(x+\zeta) = \Lambda_k(x)$ for every k, so the evidence constraints, the weights, the information acquired, and the act are all unchanged.

(ii) Otherwise ζ moves Λ_k only for the alternatives k that read it, through their statistics s_k, and moves the bound of any other alternative only through the denominator of (5)*.*

The proof is immediate from (13): every term of $\log\Lambda_k$ depends on x only through the statistics s_j, and a component no model reads leaves each of them unchanged, while a component only an alternative reads leaves s_0, and so the default's surprise, unchanged. A cue the default does read is ordinary evidence under the default and moves every bound through its surprise. In associative-memory models, adding a payoff-irrelevant cue changes which representation of the situation comes to mind and so changes behavior (Bordalo et al., 2026). Here a cue sways behavior only if it is payoff-relevant under some mental model on the menu and lifts that model over the activation threshold of Proposition 1; if it does, the curiosity this triggers tests the shift, as the next result shows.

Corollary 3 (Exposure and the path of the shift)**.** *Let observations along the direction she chooses be generated by a density f, and let f_k and f_0 be the predictive densities of alternative k and the default along it. Each observation changes $\log\Lambda_k$ in expectation by $\mathrm{KL}(f\,\|\,f_0) - \mathrm{KL}(f\,\|\,f_k)$: the shift continues on the DM's own exposure if the alternative fits the observations better than the default, and reverses otherwise. If $f = f_k$, the change is $\mathrm{KL}(f_k\,\|\,f_0)$, positive whenever the two predictive distributions differ; if $f = f_0$, it is $-\mathrm{KL}(f_0\,\|\,f_k)$. At $w_k = 0$ no information about the alternative is sought and neither occurs.*

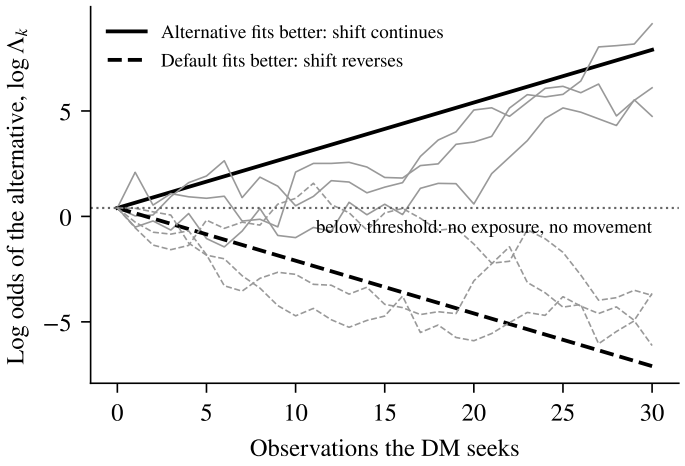


FIGURE 4. THE PATH OF A SHIFT UNDER ENDOGENOUS EXPOSURE

Notes: The figure plots the log odds of an activated alternative against the number of observations the DM seeks along her chosen direction (Corollary 3), with Gaussian predictive densities of equal variance whose means differ by $1/\sqrt{2} \approx 0.71$ standard deviations, so that each observation moves the log odds by ± 0.25 in expectation; thick lines are expected paths and gray lines are simulated ones. Below the threshold the DM seeks no such observations and the log odds stay put.

This delivers persistence from choices rather than from a state variable: a shift that crosses the threshold tends to carry itself forward when the alternative fits the observations better than the default and to correct itself otherwise, because the DM now looks where the answer lies, while below the threshold her exposure stays governed by the default and nothing she sees moves the bound (Figure 4).

Remark 1 (Misdirected search). When one of the two mental models is true, reversal of a mistaken shift does not require the DM to look along d^*. If the default is true, an observation along any direction d changes the log Bayes factor of the alternative in expectation by $-\mathrm{KL}(f_0 \,\|\, f_k)$ along d, strictly negative whenever the two predictive distributions differ, and by Lemma OA.2(i) the odds of the alternative cannot exceed its updated Bayes factor. The argument holds for any rule that selects d, so a tendency to seek confirming evidence slows correction, and stalls it only along directions on which the models make identical predictions, but cannot reverse it. If neither model is true, the sign of $\mathrm{KL}(f \,\|\, f_0) - \mathrm{KL}(f \,\|\, f_k)$ can depend on d, and misdirected search can select directions along which the alternative fits better and so sustain the shift.

Corollary 2 separates two questions that associative-memory models answer together. Whether context moves behavior depends on the menu of mental models: some model must read the cue. Whether the movement lasts depends on the information environment and the menu. Within the model, Corollary 3 and Remark 1 exhaust the ways a context effect can persist: evidence that is scarce, too costly to seek given the weight the alternative carries (Corollary 1), or biased toward the mistaken model, or a menu on which no mental model is true. The list is derived in the linear-Gaussian setting with one alternative held; outside it, it is the hypothesis the model offers for where to look when a context effect does not fade. Because the DM has an endogenous drive to seek this evidence herself, policy can make it accessible instead of delivering it: for farmers who distrust fertilizer that looks clumped but meets nutrient standards (Michelson et al., 2021), the remedy is quality testing they can obtain on their own.

4.2 Interference effects

In psychology, interference is the finding that similar memories compete at retrieval, so that recalling one item makes related items harder to recall (Underwood, 1957; Anderson and Neely, 1996; Anderson, 2003; Kahana, 2012). In economic associative-memory models, different experiences compete for retrieval, so recalling one inhibits recall of others and a prime can crowd out better information (Bordalo et al., 2023, 2025). For

example, in Amazon's early days, people who thought of bookstores saw it as a firm in a mature sector: these frequent experiences crowded out the fewer but more relevant experiences of disruptive technologies, which would have suggested a future giant (Bordalo et al., 2025). In these models, the representation with the greatest cue similarity is the one that comes to mind.

By contrast, in this framework the menu delivers a result that no comparison of one alternative with the default can. When each of several alternatives is worth more than the default, none may be activated, because the same evidence must justify each of them and justifies none enough: the automatic matching of cues against stored mental models weighs all of them at once (Section 2.2.3), so the evidence cap of each alternative is deflated by every other that looks plausible, and the cost function penalizes spreading attention weights across them (Lemma 2). These are interference effects, but every alternative remains available; rivals interfere by dividing the support the evidence provides. What interferes is a feature of the information environment acting on a given menu, not a failure of retrieval.

Before turning to how the caps interact, it helps to know what the solution with many alternatives looks like. It is not a portfolio.

Proposition 8 (Many alternatives)**.** *Let φ be linear and fix n. At any optimum, every alternative but at most one is either not held, $w_k^* = 0$, or held at its evidence cap, $w_k^* = \overline{W}_k(x)$. In particular, if no cap binds at the optimum, at most one alternative is held.*

Holding a fixed total weight off the default, the objective is convex in how that weight is split: the value of the single act is a maximum of functions linear in the weights, the hedging cost is a variance and so concave, and the cognitive cost penalizes spreading (Lemma 2). A convex objective is maximized at a corner of the feasible set, and the corners of a simplex cut by caps have at most one coordinate strictly between zero and its cap. The DM therefore fills alternatives to their caps and deliberates over at most one. This is why the single-alternative analysis of Section 3 is a reduction rather than a special case: the alternatives interact through the caps and, when held together, through the hedging costs and the costs of attention weights they share. Proposition 8 needs only that the cost of attention weights be strictly concave along splits of a fixed total weight, of which $\eta > \gamma$ is the quadratic case; the crowding results below need no assumption on costs and no symmetry across alternatives.

Proposition 9 (Fragmentation)**.** *Fix x and let $\overline{W}_k$ be given by* (5)*.*

(i) (Crowding.) *For each $k \geq 1$, $\overline{W}_k$ is strictly decreasing in L_j for every $j \neq k$ with $j \geq 1$. In particular $\overline{W}_k < L_k/(L_0 + L_k)$, the bound that would obtain were k the only alternative on the menu.*

(ii) (Breadth dilutes.) *Let a treatment scale the likelihoods of a set $J \subseteq \{1, \ldots, K\}$ by a common factor e^{β}, $\beta > 0$. For $k \in J$, the resulting bound is strictly smaller when $|J| \geq 2$ than when $J = \{k\}$.*

(iii) (Joint failure.) *Consequently, in the all-or-nothing regime, a feature vector that raises several L_k moderately can leave every $\overline{W}_k$ below the level alternative k needs in Proposition 1 at every n, not only at $n = 0$. If, in addition, no two alternatives are worth holding together, $(m_j - m_0)(m_k - m_0)/\lambda \leq \eta$ for every pair, then $w_k^* = 0$ for all $k \geq 1$ and, under Assumption 1, $n^* = 0$. In the graded regime crowding caps each weight at $\overline{W}_k$, and no information is acquired once $\sum_{k \geq 1} \overline{W}_k / \underline{w}_k \leq 1$, with $\underline{w}_k$ the level in Corollary 1.*

Proposition 9 is the interference effect in its basic form. The signal's diagnostic power, large in aggregate, is fragmented across the menu, and the default remains defensible: fragmentation can produce a curiosity failure even when every alternative outperforms the default. Where crowding lowers the bounds without pushing them below threshold, alternatives are activated at smaller weights and, by Proposition 2, investigated less, so the joint failure of part (iii) is the limit of a graded effect rather than a knife edge. Joint failure needs the pairwise condition because alternatives that agree with one another can be pooled: holding two together saves hedging when they pull the action the same way, and if the saving exceeds the extra cost of attention

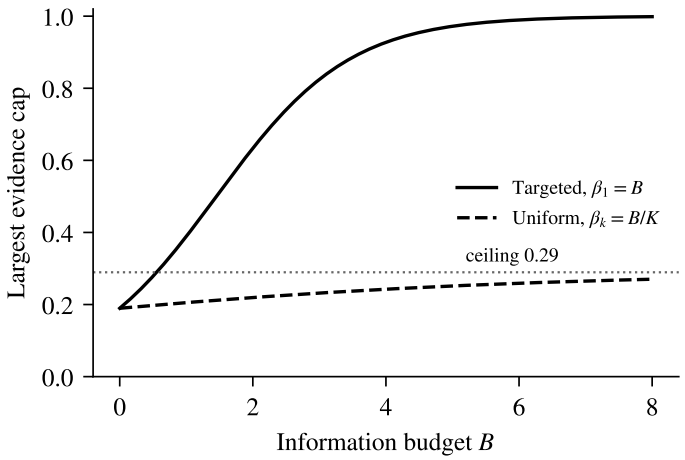


FIGURE 5. TARGETING VERSUS BREADTH IN INFORMATION INTERVENTIONS

Notes: The figure plots the largest evidence cap among $K = 4$ alternatives against an information budget B that multiplies each likelihood L_k by e^{β_k}, with $(L_0; L_1, \ldots, L_4) = (1.0; 0.55, 0.50, 0.45, 0.40)$. The targeted treatment puts the whole budget on the most likely alternative ($\beta_1 = B$); the uniform treatment spreads it evenly ($\beta_k = B/K$). The dotted line is the ceiling $\max_k L_k / \sum_j L_j \approx 0.29$ that no uniform budget exceeds (Proposition OA.5 in the Online Appendix).

weights, their bounds effectively add up (the stacking premium of Proposition OA.6 in the Online Appendix is then negative). Crowding therefore stifles exploration most when the rival interpretations disagree about what to do. Curiosity is therefore non-monotone in the number of rivals: zero with none, positive with one that is activated, and zero again once enough rivals fit the evidence, provided spreading attention weights across them is costly enough, whether or not the rivals are symmetric (Corollary OA.2 in the Online Appendix).

Corollary 4 (Crowding by unpromising rivals)**.** *Fix x and let alternative j satisfy $\Delta U_j < \min\{D_j, \gamma/2\}$, so that by Proposition 1 it is never activated on its own, whatever its bound. Without the evidence constraint, adding j to the menu cannot leave the DM with the default alone if she held an alternative before. With it, adding j multiplies every other bound by $(1+\sum_{i\neq j}\Lambda_i)/(1+\sum_i \Lambda_i) < 1$, so it can prevent the activation of an alternative k that would otherwise be activated, leave the DM with the default alone, and, under Assumption 1, set $n^* = 0$.*

Without the constraint, adding j only enlarges the set of feasible weights, and holding j alone is never worthwhile, so whatever she explored before she still explores. With it, j need receive no weight, yet it enters the denominator of (5) through Λ_j alone, whatever its value. The DM can ignore an unpromising interpretation but not the evidence for it: a plausible story she would never entertain still crowds out one she would. In Bayesian model averaging, by contrast, a plausible rival crowds out others only by receiving posterior weight itself; here it receives none, and it affects exploration through its likelihood, not through the attention weight it receives. Under optimal expectations such an interpretation could simply be dropped from the menu; here it cannot, for the reason given in Section 2.2.3: dropping it because it is unattractive is the wishful move the constraint forbids, and dropping it because it is implausible requires evidence against it, which is what critical thinking does in Section 6 and what refuting rivals one at a time does (Corollary OA.3 in the Online Appendix). The claim is testable, and it separates the framework from accounts based on retrieval or limited attention: adding a plausible but unattractive rival should lower the weight a person places on a promising alternative even though the rival receives no attention weight, with an effect that scales with the rival's plausibility rather than its appeal, and that disappears once evidence rules the rival out, although evidence about the promising alternative itself is unchanged. The same holds for a contextual cue that only such a rival reads, which lowers the promising alternative's bound (Corollary 2(ii)) although the alternative it favors is never entertained.

Interference effects have two direct applications to the design of information interventions.

Application 1: Targeting versus breadth. Consider an information intervention with a fixed amount of evidence to distribute across alternatives, so that each alternative's likelihood rises with the share it receives. Concentrating the budget on the most likely alternative drives its evidence cap to one as the budget grows, whereas spreading it uniformly leaves every cap below $\max_k L_k / \sum_{j \geq 1} L_j < 1$ at every budget, however large (Proposition OA.5 in the Online Appendix). Breadth fails not because it delivers less information but because it supports competing interpretations at once, raising every likelihood in the denominator of every bound, so the ratios barely move. Figure 5 shows the two paths for $K = 4$: the targeted bound passes 0.9 at a budget of about four, while the uniform bound stays below 0.29 at any budget. What stifles exploration is thus not the number of reasons offered for a program but their incoherence: separate interpretations that compete for the same evidence make several moderately plausible and none plausible enough, whereas reasons that cohere into one interpretation are one mental model and do not divide the evidence. A program is therefore better promoted through a coherent presentation.

Application 2: Sequencing versus breadth. Suppose two alternatives must eventually be entertained. Under (3) the cost of holding both at weights w_1, w_2 is $\frac{\gamma}{2}(w_1^2 + w_2^2) + \eta w_1 w_2$, whereas two separate activation–signal–resolution cycles cost $\frac{\gamma}{2}(w_1^2 + w_2^2)$ in total: holding them simultaneously costs a stacking premium of $\eta w_1 w_2$ and raises the hedging cost when the two alternatives pull the action in opposite directions, so absent discounting sequencing dominates stacking unless they point the same way and spreading attention weights is cheap (Proposition OA.6 in the Online Appendix). Sequencing also arises on its own. If one alternative clears its threshold, it is investigated first; if the evidence goes against it, every other bound rises, so a second may clear its threshold in turn. If none is worth holding, alone or jointly, no bound moves and the joint failure persists until a new exogenous observation arrives (Corollary OA.3 in the Online Appendix). The DM can thus climb out of a fragmented menu on her own, eliminating hypotheses one at a time, but only after something has lifted one alternative over its threshold.

5 Creative Thinking: Compressing the Space of Mental Models

Sections 2–4 take the menu of mental models as fixed, and with a fixed menu the model is incomplete in two ways. First, curiosity resolves itself: the evidence it gathers settles the model uncertainty that prompted it (Proposition 4), so without a source of new alternatives exploration dies out. Second, many practical situations involve many dimensions at once: a farmer chooses among many inputs and practices whose returns depend on one another, so the combinations to consider multiply (Laajaj and Macours, 2026). A DM who represented every possibility as a separate mental model would rarely explore at all, because of the interference effect of Section 4.2. Together these raise a question that a fixed menu cannot answer: how does a mind with limited capacity replenish its menu while keeping it small enough to explore at all?

This section and the next relax the fixed menu and model creative and critical thinking as two forms of *cognitive dimension reduction*. Exploration requires describing many possible mental models with few reusable parts, and comparing few of them at a time. Creative thinking does this by compression, building new mental models from parts the DM already holds and pooling situations under common parameters; critical thinking does it by pruning, keeping chance fits off the menu. I begin by defining templates—the reusable parts of mental models.

5.1 Templates and mental models

Let ℓ index situations, such as plots, lessons, or problems, and let $\mathscr{T} = \{1, \ldots, M\}$ be a library of templates. Informally, a template is a reusable piece of explanation: an underlying cause together with the observable

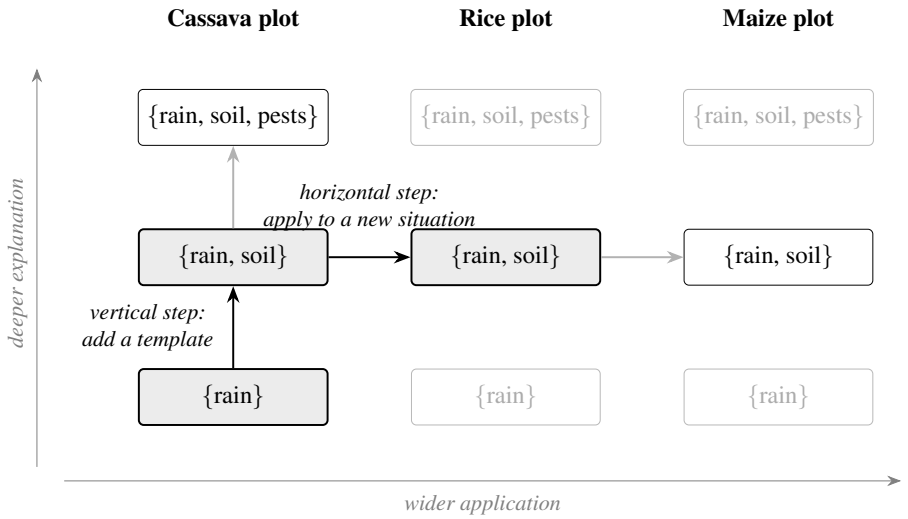


FIGURE 6. VERTICAL AND HORIZONTAL CREATIVE THINKING

Notes: Each box is a node, a set of templates applied to a situation; shaded boxes are nodes the DM holds, gray boxes nodes she could construct. A vertical step adds a template in a given situation, here crediting a cassava harvest to the soil as well as the rain (Proposition 10); a horizontal step applies the node to a new situation, here the rice plot (Proposition 11). If the soil template is invariant, a vertical step that adds it extends how far the node transfers (Proposition 12).

features it produces, such as rain, which shows up in wet soil and high yields, or pests, which show up in damaged leaves and poor yields. A farmer who credits a neighbor's harvest to both the rain and the soil has combined two templates (Behrens et al., 2018; Whittington et al., 2020).[27]

Creative thinking works with these parts in two ways. A farmer who credits her cassava harvest to the rain can add the soil to that explanation, a *vertical* step, or carry the explanation that works on her cassava plot over to her rice plot, a *horizontal* step (Figure 6). If each mental model were a separate hypothesis, as in Sections 2–4, the mental models these steps produce could at best be listed in advance, and nothing would connect a mental model to the one it extends or to the situation it is carried from. Representing the steps themselves requires building mental models from reusable parts.

Formally, a template represents a causal mechanism as the DM understands it: a latent cause together with the edges from it to the observable features it moves, as in a structural causal model (Pearl, 2009). The causal reading belongs to her mental models; the statistical representation does not by itself identify causal mechanisms. What the DM observes is then the sum of the effects of the causes she credits, $x = \sum_{t \in S} \vartheta_t b_t + \varepsilon$, where S is the set of templates she credits and ϑ_t is how strongly cause t acts. This is a factor representation, but the factors are explanatory mechanisms she holds, not latent traits to be measured as in factor analysis. I take the linear Gaussian case so that the rules for admitting a new mental model have closed forms. A template is then a tuple $\tau_t = (b_t, \mu_t, \sigma_t^2, \rho_t)$. The loading vector $b_t \in \mathbb{R}^d$ records which observable features the cause moves: the direction in feature space along which it shows up, so that for rain it loads on soil moisture and yields. The prior $\mathcal{N}(\mu_t, \sigma_t^2)$ records how strongly the cause is expected to act, and its variance is the template's vagueness; a template the DM has not yet used is centered, $\mu_t = 0$, since she has no view of how strongly it acts. The *generality* $\rho_t \in [0,1]$ sets how far that strength is invariant across situations. A *node* is a

[27] Templates come from slower learning or from innate structure (Spelke and Kinzler, 2007), and recent work suggests what one looks like in the brain: neural populations in the hippocampus and prefrontal cortex encode what they learn on low-dimensional surfaces, or manifolds, with a shared geometry, so that a pattern learned in one setting carries over to new ones (Bernardi et al., 2020), and the hippocampal manifold expands as experience accumulates (Zhang et al., 2023).

set of templates $S \subseteq \mathscr{T}$ applied to one or more situations. Its generative process in a situation ℓ is

$$x_\ell \mid \vartheta \;\sim\; \mathscr{N}\Big(\sum_{t\in S} \vartheta_{t,\ell}\, b_t,\; \Sigma\Big), \qquad \vartheta_{t,\ell} = \mu_t + \omega_t + e_{t,\ell}, \qquad \omega_t \sim \mathscr{N}(0, \rho_t \sigma_t^2), \qquad e_{t,\ell} \sim \mathscr{N}\big(0, (1-\rho_t)\sigma_t^2\big), \tag{15}$$

with all ω_t and $e_{t,\ell}$ independent of one another and of the noise. Templates contribute additively to what is observed, as causes do in a linear structural model.

Each situation's parameter has the template's prior whatever its generality, and any two situations' parameters have correlation ρ_t, so a broader claim is not also a vaguer one. A template with $\rho_t = 0$ makes no claim across situations, since its parameters are independent draws from the prior. A template with $\rho_t = 1$ claims that one parameter governs every situation the node is applied to, and the set of situations over which its parameter is shared in this way is its *reach*: the situations across which the mechanism is invariant (Peters, Bühlmann, and Meinshausen, 2016), so that it transports from one to another (Pearl and Bareinboim, 2014). Intermediate values of ρ_t express partial pooling: the DM is unsure how far the mechanism transports.

Templates are not a new primitive but a factorization of the one the paper already has.

Lemma 3 (Mental models are nodes)**.** *With one template t and one situation,* (15) *is the generative process* (1) *of Section 2.2.1 with $b_k = b_t$, $\mu_k = \mu_t$, and $\sigma_k^2 = \sigma_t^2$. For any node S in any situation, the marginal likelihood of x, and hence the evidence cap* (5)*, the costs* (3) *and* (7)*, and the objective* (4)*, are those of Section 2 with the node in the place of the mental model.*

The extension departs from the baseline in two ways only: a node may combine several templates, and a template's parameter may be shared across situations. The correspondence is otherwise exact, object by object:

Section 2	Sections 5–6	How the second generalizes the first
mental model k	node S	a node with one template in one situation is a mental model
loading b_k	loading matrix B_S	B_S stacks one loading b_t per template; with one template, $B_S = b_k$
prior (μ_k, σ_k^2)	template priors (μ_t, σ_t^2), generality ρ_t	each template keeps its own prior; ρ_t sets how far its parameter is shared across situations
initial signal x	residual $r = x - B_S\mu_S$	the initial signal net of what the held node predicts; with no node held, $r = x$
evidence cap $\overline{W}_k(x)$	cap of the new node	equation (5) with nodes in place of mental models
activation of k	admission of a step	the new node's cap in the comparison with the old exceeds one half, that is, its Bayes factor against the old node exceeds one

A node is thus a causal graph, as in models of causal narratives (Spiegler, 2016; Eliaz and Spiegler, 2020), assembled from reusable mechanisms.

Definition 2 (Two types of creative thinking)**.** *Vertical creative thinking* adds a template to a node, $S \mapsto S \cup \{t\}$, in a given situation. *Horizontal creative thinking* applies a node to a new situation, for given templates.

In causal terms, a vertical step adds a cause to the graph, and a horizontal step transports the graph to a new situation. Judging such steps by Bayes factors with their Occam penalties is how cognitive science models the discovery of structure (Kemp and Tenenbaum, 2008; Griffiths and Tenenbaum, 2009; Tenenbaum et al., 2011); what the framework adds is the economic layer: the cost of attempting a step, the evidence cap on the new node, and the thresholds these produce.[28]

[28] Imagining a new experience depends on the memory system and works by recombining stored experience (Hassabis et al., 2007;

The two steps reduce dimension in different directions.[29] A vertical step is judged as a comparison between two nodes, so the evidence is not divided among every assembly the library could form; a horizontal step pools situations, in the limit replacing a free parameter per situation with a single shared one.

Attempting a step to a node S' costs mental effort F, and the DM attempts it when the value of holding S', at the weight its evidence cap admits, covers the cost:

$$V\left(\overline{W}_{S'}(x), n^*\right) - V\left(\overline{W}_S(x), n^*\right) \;\geq\; F. \tag{16}$$

A step to S' is *admitted* if the evidence favors S' over the node it would replace, that is, if the evidence cap of S' in a comparison of the two exceeds one half.

5.2 Vertical creative thinking

Fix one situation and a node S. Write B_S for the matrix of the node's loadings, Σ_S for the diagonal matrix of their prior variances, and

$$V_S := \Sigma + B_S \Sigma_S B_S'$$

for the node's predictive covariance of the features. Let $r := x - B_S\mu_S$ be the part of the initial signal that node S did not predict, and let $h_t := b_t' V_S^{-1} b_t$ be the precision with which template t's direction can be read against everything node S already explains.

Proposition 10 (When a vertical step is admitted)**.** *Let $\upsilon_t := \sigma_t^2 h_t$ measure the template's vagueness relative to the precision with which its direction can be read, the counterpart of σ_k^2/ν_k^2 in Proposition 3. The log Bayes factor of the node $S \cup \{t\}$ against S is*

$$\log \Lambda_{S\cup t:S} \;=\; \tfrac{1}{2}\Big[\,\frac{\upsilon_t}{1+\upsilon_t}\, z_t^2 - \log(1+\upsilon_t)\Big], \qquad z_t := \frac{b_t' V_S^{-1} r}{\sqrt{h_t}}, \tag{17}$$

where z_t is the standardized surprise of the residual along the template's direction, the surprise channel of Proposition 3 in one direction, and $\log(1+\upsilon_t)$ is the template's Occam factor. In a comparison of the two nodes, the step is admitted, $\overline{W}_{S\cup t} > 1/2$, if and only if

$$z_t^2 \;>\; \frac{1+\upsilon_t}{\upsilon_t}\log(1+\upsilon_t), \tag{18}$$

a threshold that is increasing in υ_t and tends to one as $\upsilon_t \to 0$.

Proposition 10 shows where compression pays: the comparison is between two nodes, so the test is one-dimensional, the surprise of the residual along the template's direction net of an Occam penalty for its vagueness. Surprise and encoding both enter through z_t, the second through features S ignores, which cannot move any decision made under S but can justify a new template (Corollary OA.4 in the Online Appendix), while specificity is the Occam penalty, so a vaguer template is harder to adopt, not easier. Compression has a price. If the DM tries many templates after seeing x and keeps the best, the candidates she tried form a menu in all but name, and the crowding that would have disciplined that menu is lost.

Schacter and Addis, 2007; Gupta et al., 2010). The DM chooses steps rather than finished creations, since a step can be ranked before the creation is conceived; preferences decide whether a step is attempted, surprise gives it a direction, and evidence decides whether it is admitted (Campbell, 1960). Vertical steps are abductive, horizontal steps analogical. The approach is close to combinational creativity (Koestler, 1964; Boden, 2004; Knoblich et al., 1999), with M templates describing up to $2^M - 1$ mental models (Weitzman, 1998), and differs from models of growing awareness, in which unforeseen options arrive from outside (Karni and Vierø, 2013); transformational creativity (Boden, 2004) is read here as the slower learning by which new templates form, which lies outside the model.

[29] They are the knowledge counterparts of the two kinds of innovation in endogenous growth: a vertical step is a step up a quality ladder (Aghion and Howitt, 1992), a horizontal step an expanding variety (Romer, 1990).

5.3 Horizontal creative thinking

Now hold the templates fixed and ask whether a node applies across N situations. Consider one template, with prior $\mathcal{N}(\mu, \sigma^2)$ and generality ρ; in a node with several templates the argument applies to each template's parameter given the others. Let s_ℓ be the template's sufficient statistic in situation ℓ, read as in Lemma 1, and measure it in units of its reading noise, so that $s_\ell \mid \vartheta_\ell \sim \mathcal{N}(\vartheta_\ell, 1)$. The *local* node sets $\rho = 0$ and treats the N parameters as unrelated draws from the prior. The *general* node shares the parameter across situations; I state the invariant case $\rho = 1$, in which one parameter governs all N situations, and leave intermediate ρ to the Online Appendix. Write $d_\ell := s_\ell - \mu$ for each situation's departure from the prior, $\bar{d} := N^{-1}\sum_\ell d_\ell$ for the departure the situations share, and $SS := \sum_\ell (d_\ell - \bar{d})^2$ for their dispersion.

Proposition 11 (When a horizontal step is admitted)**.** *At $\rho = 1$, the log Bayes factor of the general node against the local node is*

$$\log \Lambda_{\text{gen:loc}} = \underbrace{\tfrac{1}{2}\Big[N\log(1+\sigma^2) - \log(1+N\sigma^2)\Big]}_{\textit{Occam gain}} - \underbrace{\tfrac{1}{2}\frac{\sigma^2}{1+\sigma^2}\,SS}_{\textit{dispersion penalty}} + \underbrace{\frac{N(N-1)\,\sigma^2}{2(1+\sigma^2)(1+N\sigma^2)}\,\bar{d}^2}_{\textit{shared-departure gain}}. \tag{19}$$

It is zero when $N = 1$. For $N \geq 2$, it is increasing in the shared departure $\bar{d}^2$ and decreasing in the dispersion SS, so the general node is admitted if and only if SS falls below a threshold that is increasing in $\bar{d}^2$. The Occam gain is increasing in N.

For $0 < \rho < 1$ the same three terms appear with smaller Occam gain and smaller dispersion penalty, both increasing in ρ (Proposition OA.7 in the Online Appendix).

Horizontal creative thinking reduces dimension across situations: the local node carries N free parameters, the general node one, and evidence from different situations then bears on a single parameter rather than being divided among N, the cross-situation counterpart of concentrating information on one alternative (Section 4.2). The dispersion penalty gives a reading of a classic finding on social learning: at the onset of the Indian Green Revolution, farmers learned from neighbors' yields for wheat but not for rice, whose yields were sensitive to unobserved soil characteristics, and rice growers experimented on their own land instead (Munshi, 2004). Where situations differ, the step that carries a neighbor's experience to one's own plot is not admitted, and evidence must be gathered within the situation. Transfer can fail across domains as well as across plots. Banerjee et al. (2025) find that children who work in Indian markets calculate fluently in market transactions yet fail simpler abstract problems of the kind taught in school, while schooled children show the reverse pattern. In the framework, the arithmetic each group has learned is a node bound to the situations in which it was acquired; without a template shared across the two settings, the horizontal step is not admitted and the knowledge stays where it was formed.

5.4 Deeper explanations apply more widely

Every template is a mechanism; an *invariant template* is one whose mechanism acts with the same strength in every situation in which it operates: its generality is $\rho_t = 1$ on its reach R_t, and outside R_t it does not contribute to what is observed.

Proposition 12 (Invariant templates determine the extent of horizontal transfer)**.** *Let $S' = S \cup \{t\}$ add an invariant template t whose parameter is common to a set of situations R_t. Then:*

(i) The reach of S' is R_t: the parameter of t is shared on R_t, and outside R_t the node is as before, so the reach is a property of the template rather than a choice.

(ii) For $N \geq 2$ situations in R_t whose evidence agrees ($SS = \bar{d} = 0$), the transfer Bayes factor of Proposition 11 is strictly larger under S' than under any node with $\rho < 1$.

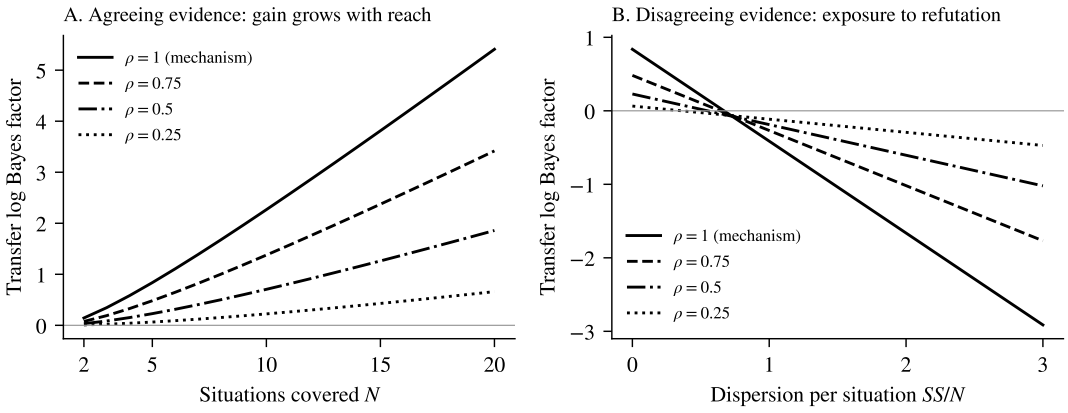


FIGURE 7. DEEPER EXPLANATIONS APPLY MORE WIDELY

Notes: Both panels plot the log Bayes factor of the general node against the local node from the general form of Proposition 11 (Online Appendix), with $v = \sigma^2 = 1$ and no shared departure ($\bar{d} = 0$), for four levels of generality ρ. Panel A varies the number of situations the template covers when their evidence agrees ($SS = 0$). Panel B fixes $N = 5$ and varies the dispersion of the evidence across situations. An invariant template ($\rho = 1$) gains most from agreement and loses most from disagreement.

(iii) For situations whose evidence disagrees, the penalty on dispersion is increasing in ρ and largest at $\rho = 1$.

Invariant templates are the limit of compression: one parameter for every situation in which the mechanism operates, with a reach set by the template rather than chosen. The property that lets agreeing evidence pool (part (ii)) also makes disagreeing evidence count most heavily against the template (part (iii)). Figure 7 shows the two faces: with agreeing evidence the gain from an invariant template grows fastest in the number of situations covered, and with disagreeing evidence it falls fastest. Maximal compression is maximal exposure to refutation, which is why compression needs the discipline of the next section. The DM chooses whether to apply a node to a new situation; whether it transfers, and how far, is decided by the templates it contains. Vertical and horizontal creative thinking are therefore not independent, yet neither substitutes for the other: the vertical step sets how far a node can transfer, and the horizontal step is still required to realize the transfer and must be justified by the evidence in each situation.[30] Because creation draws only on the templates the DM holds, creative thinking is not a fixed trait but a capability that can be cultivated: providing templates, and above all invariant ones, enlarges both the set of mental models she can construct and the range of situations to which they transfer.

This matters for training teachers and agricultural extension officers. Practice with a set protocol leaves them with a precise local mental model, which pays a small Occam penalty and is easily activated where it was designed but carries no claim beyond it; understanding why a practice works supplies an invariant template, whose gain is zero in a single setting and grows with the number of classrooms and farms it covers (Proposition 11), at the cost of greater exposure to evidence that disagrees (Proposition 12(iii)). This qualifies a common presumption that teaching why a practice works is enough to make trainees generalize: providing a generalizable explanation is not the same as developing the ability to generalize. The template does not deliver the adaptations ready-made: each is a horizontal step the trainee must take and test herself, and it happens only if her curiosity clears the threshold in the new setting (Section 3.1), so training in mechanisms pays through her continued curiosity and horizontal creative thinking, and through the opportunities that make each application pay, not in place of them. This complements recent work that formalizes how teaching mechanisms improves decisions in agriculture: in Davies and Sankar (2026) and Sankar et al. (2026), a mechanism is a known truth

[30] In the language of growth theory, an invariant template is a general-purpose technology (Helpman and Trajtenberg, 1998): one step up the ladder raises explanatory quality wherever the mechanism operates and, by part (ii), the return to expanding variety over its reach. Models of innovation treat the two kinds of step as separate activities (Aghion and Howitt, 1992; Howitt, 1999).

that lets a Bayesian design better signals and generalize within a single decision; here it is a hypothesis whose value is realized only across situations, and only if the horizontal step is taken and admitted.[31]

The framework also offers an interpretation of skill. Labor economics typically treats skill as a stock that raises productivity or as the capacity to perform particular tasks (Becker, 1964; Acemoglu and Autor, 2011), with growing returns to nonroutine tasks that require solving unfamiliar problems (Autor et al., 2003). Here skill is the library of templates a person holds. Templates suited to a problem let a DM compress the same evidence further and activate the right mental model with fewer observations, so she adapts faster to situations she has not seen before. More templates, however, also supply more candidate interpretations that compete for the same evidence, so skill lies in holding the right templates rather than many. The advantage should be largest where problems are new and involve many interacting inputs. In agriculture, for example, Laajaj and Macours (2026) find that high-skill farmers learn more from agronomic trials, and model skill as the precision with which a farmer reads each observation; the framework predicts their finding that the advantage concentrates in decisions with many interacting inputs.

6 Critical Thinking: Pruning the Menu with Fresh Evidence

Compression creates its own risk. A library of templates makes candidate mental models cheap to construct, and a DM who searches them after seeing the evidence and keeps the best fit has considered a large menu while counting only the winner. The rivals she discarded drop out of the evidence cap, and with them the crowding of Proposition 9 that a menu of that size would have faced. Critical thinking prunes the menu, admitting only mental models that survive evidence not used to form them. That searching many explanations yields false ones is common sense; the results below say when discipline is enough, a threshold that rises with the templates' vagueness, what it buys beyond avoided errors, a higher evidence cap for the right alternative, and what it costs, a curiosity failure for correct insights.

The problem is familiar to empirical economists as multiple hypothesis testing, for which multiple-inference corrections and pre-analysis plans are now standard (Simmons et al., 2011; Gelman and Loken, 2014; Anderson, 2008; Olken, 2015; List, Shaikh, and Xu, 2019). A common cognitive failure has the same form: picking an explanation because it fits and then counting that same fit as evidence that it is right, so that the data are used twice. A farmer whose neighbor harvests far more than usual runs through possible reasons, such as the seed, the rain, the planting date, and the spacing, and settles on spacing because it matches what she saw most closely; with enough candidate reasons, one is bound to match by chance, so the match alone says little. Reading meaning into chance coincidences is also how superstitions form (Skinner, 1948). I model one core component of critical thinking, the discipline against after-the-fact fitting: the disposition to distinguish predictions from after-the-fact explanations, so that every mental model, new or old, is judged only by evidence that was not used to form it. It is close to the disposition to evaluate arguments independently of prior belief (Stanovich and West, 1997) and opposite to confirmation bias (Nickerson, 1998).[32]

[31] Davies and Sankar (2026) show that conceptual knowledge is valuable because it lets the agent design more informative signals, a value that vanishes as signals become abundant; Sankar et al. (2026) find that adding mechanistic explanations to fertilizer demonstrations in Uganda helps farmers generalize across soils and raises yields. Here the value of a mechanism arises through transfer across situations rather than sampling within one, and can fail: a mechanism can go unactivated, be crowded out by rivals, or be refuted.

[32] Logic is not absent but assumed error-free: deriving what a mental model predicts from its templates is deduction, which the generative structure performs for the DM, so critical thinking here is the empirical half of hypothetico-deductive reasoning, testing those predictions against evidence not used to form them. The analysis of arguments on their internal merits and the evaluation of sources (Ennis, 1987; Halpern, 1998) lie outside the model.

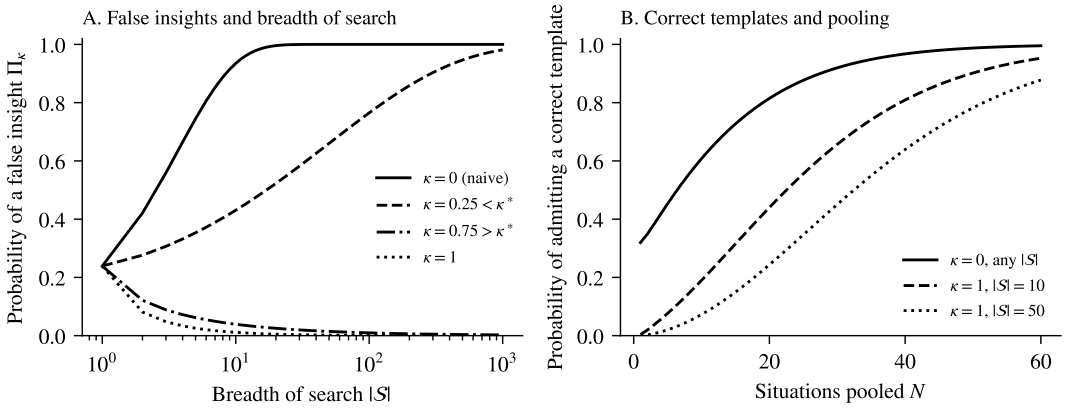


FIGURE 8. CRITICAL THINKING, BREADTH OF SEARCH, AND POOLING

Notes: Both panels set $\upsilon = 1$, so the threshold in Proposition 13 is $\kappa^* = 1/2$. Panel A plots the probability of a false insight $\Pi_\kappa(|\mathscr{S}|)$ against the breadth of search for four levels of critical thinking, with the candidates independent under S. Panel B plots the probability that a correct template is admitted against the number of situations N whose evidence bears on it (Corollary 5), with $h_t \vartheta_t^2 = 0.36$ in each situation; the discount leaves the naive DM's bar unchanged, so her curve does not depend on $|\mathscr{S}|$.

6.1 The search discount

Fix a node S and an observation x. After seeing x the DM searches a class $\mathscr{S} \subseteq \mathscr{T}$ of candidate templates and keeps $t^* := \arg\max_{t \in \mathscr{S}} \Lambda_{S \cup t:S}(x)$, with the Bayes factor of Proposition 10. The discount she should apply is not a free choice; it follows from the evidence constraint. Searching $\mathscr{S}$ puts the $|\mathscr{S}|$ candidate nodes $S \cup \{t\}$ on the menu alongside S. The search asks one question, whether some extension of S explains x, and answers it in $|\mathscr{S}|$ ways. I assume the automatic assignment of Section 2.2.3 gives that question the prior weight it gives a single alternative named in advance, equal to that of S, and spreads it evenly over the candidates; with $|\mathscr{S}| = 1$ this is the uniform prior of (5). The evidence cap of the winner in its comparison with S is then $\Lambda_{S \cup t^*:S}/|\mathscr{S}|$ in odds, because the winner carries only its share $1/|\mathscr{S}|$ of that weight. The comparison with S alone is the lenient one: the other candidates would lower the winner's cap further through the denominator. The log odds the constraint allows are therefore $\log \Lambda_{S \cup t^*:S}(x) - \log |\mathscr{S}|$: the penalty is the uniform prior over the searched candidates, written in logs. The uniform prior is a convenience, not a restriction: with unequal prior weights π_t, $\log |\mathscr{S}|$ becomes $\log(1/\pi_{t^*})$, the effective breadth of the search, and what matters below is only that the penalty grows logarithmically in the number of candidates, which it does for any prior whose weights stay within a bounded factor of one another. Critical thinking is the degree to which deliberation registers this penalty, in the same sense in which Section 2.2.3 let it register a fraction of the evidence.

Definition 3 (Critical thinking)**.** A DM with critical thinking $\kappa \in [0,1]$ evaluates a node constructed by searching a class $\mathscr{S}$ after observing x at the log odds

$$\log \Lambda_{S \cup t^*:S}(x) \; - \; \kappa \log |\mathscr{S}|. \tag{20}$$

A model specified before x, such as one assembled from templates learned earlier, has $|\mathscr{S}| = 1$ and is not discounted, and neither is evidence that was not used in the search.

With $\kappa = 1$ the DM fully applies the search penalty that the evidence constraint implies in the comparison with S; with $\kappa < 1$ she departs from the constraint, and with $\kappa = 0$ she treats the winner as if it had been named in advance, counting the data once to find it and once to believe it. The discount depends on the breadth of the search, not on its outcome: an explanation named before the evidence is treated like a hypothesis registered in a pre-analysis plan, while one found by searching many candidates is held to a higher bar.

6.2 Search breadth and false insights

A *false insight* is a template admitted when the initial signal was generated by S. Let the candidates share the same $\upsilon_t = \upsilon$ of Proposition 10 and read directions that are orthogonal in the metric V_S^{-1}, the number of observable features growing with the search set so that such directions exist as $|\mathscr{S}|$ grows. Admission requires that the evidence cap of $S \cup \{t^*\}$ in its comparison with S exceed one half. This is a pairwise screening rule; requiring the winner's posterior on the full menu to exceed one half would be stricter, and the results below concern the pairwise rule.

Proposition 13 (Critical thinking and breadth)**.** *Under these conditions the DM admits a template if and only if the largest surprise clears the admission bar c_κ,*

$$\max_{t \in \mathscr{S}} z_t^2 \;>\; c_\kappa(\upsilon, |\mathscr{S}|) \;:=\; \frac{1+\upsilon}{\upsilon}\Big[\log(1+\upsilon) + 2\kappa \log|\mathscr{S}|\Big], \tag{21}$$

and the probability of a false insight, $\Pi_\kappa(|\mathscr{S}|)$, satisfies:

(i) If $\kappa < \upsilon/(1+\upsilon)$, then $\Pi_\kappa(|\mathscr{S}|) \to 1$ as $|\mathscr{S}| \to \infty$. In particular a naive DM forms a false insight with probability approaching one when she searches broadly.

(ii) If $\kappa > \upsilon/(1+\upsilon)$, then $\Pi_\kappa(|\mathscr{S}|) \to 0$, at a rate in $|\mathscr{S}|$ given in the Online Appendix. This part does not require independence across candidates.

(iii) For a candidate that is right, the surprise required for admission rises with the breadth of the search, by $2\kappa \frac{1+\upsilon}{\upsilon} \log|\mathscr{S}|$ in z_t^2.

The bar in (21) rises in proportion to $\log|\mathscr{S}|$, the rate at which a Bonferroni critical value for $|\mathscr{S}|$ normal tests rises, so breadth of search and discipline in judging it go together. Panel A of Figure 8 shows the divide at κ^*: below it a broad search ends in a false insight almost surely, so partial discounting cannot prevent it, and above it almost never; since κ^* rises with υ, vaguer templates demand more discipline. The threshold $\kappa^* = \upsilon/(1+\upsilon)$ is a statement about rates, not about the form of the discount: the largest of $|\mathscr{S}|$ chance surprises grows like $2\log|\mathscr{S}|$, and the discount must grow at least as fast, which it does for any prior over the candidates whose weights stay within a bounded factor of one another. Part (iii) is the seed of the potential pitfall: critical thinking also demands more of a correct insight. The discount matters beyond single steps. Each false insight a naive DM admits is a mental model that fits the evidence, so it enters the denominator of every evidence cap and dilutes the support for the alternative that is right (Proposition 9), whether or not she would ever entertain it (Corollary 4). By keeping chance fits off the menu, critical thinking can make the right alternative easier to activate. The same discount applies to old mental models: critical thinking loosens a default that owes more to after-the-fact search than the alternative does, so it is not a bias toward the familiar (Proposition OA.8 in the Online Appendix).

6.3 Fresh evidence and the pitfall of critical thinking

Fresh evidence removes the difference between the two DMs. Let $x'_1, \ldots, x'_m$ be later observations independent of x given the process that generates them, such as observations from new plots. They enter both DMs' log odds in the same way, so the critical thinker's log odds for $S \cup \{t^*\}$ against S differ from the naive DM's only by the one-time constant $\kappa \log|\mathscr{S}|$. Each fresh observation moves the log odds in expectation by how much better $S \cup \{t^*\}$ fits it than S, as in Corollary 3, and as such observations accumulate both DMs' evidence caps for $S \cup \{t^*\}$ converge to one if t^* improves the fit and to zero if it worsens it (Proposition OA.9 in the Online Appendix).

Fresh evidence plays the role of a replication or hold-out sample, and it is how the menu is pruned over time. Laajaj and Macours (2026) show what this pruning looks like in the field. After seeing agronomic trials, Kenyan farmers adopted new combinations of inputs and practices and at first made more mistakes,

pairing inputs that do not work well together; the mistakes fell as they kept experimenting on their own plots. In the language here, the trial plot is the evidence that admits a combination, the farmer's own plot is the fresh evidence that tests it, and a combination that fails the test is dropped. A node whose predictions fail on observations not used to form it sees its evidence cap fall toward zero, while the discount, charged once, matters most when evidence is scarce. Extending a mental model to a new situation, the horizontal step of Section 5, is therefore also an out-of-sample test.

Fresh evidence arrives, however, only if it is collected, and the DM collects evidence only through the mental models she holds. Say that an observation *bears on* a template t if its distribution differs under $S \cup \{t\}$ and under S. The pitfall of critical thinking is therefore a curiosity failure rather than a rejection: when such observations are sought only by a DM who places weight on $S \cup \{t\}$, the discount that keeps the template off the menu also keeps the DM from collecting the evidence that would admit it, and her log odds stay where the discount left them, whether or not the template is correct (Corollary OA.5 in the Online Appendix). The higher bar of Proposition 13(iii) is temporary only when evidence bearing on the template arrives without being sought, as when a neighbor's harvest is visible anyway or a program delivers the evidence.

The pitfall does not depend on modeling only the core component of critical thinking. Any disposition that lowers the weight on an untested interpretation lowers the evidence sought about it, since evidence is sought through the weights (Proposition 2); skepticism about sources or gaps found in an argument would add discipline of the same sign. The statement is scoped to discipline whose channel is the weight. A component of critical thinking that improves the informativeness of tests works through the precision of the signals instead, which lowers the information threshold $\underline{w}$ of Corollary 1, and so, like a narrower search or pooling across situations, offsets the pitfall; that is the same complementarity run the other way, not an exception to it. A critical thinker who tests the insights she discounts is buying evidence on a hypothesis below her information threshold, which does not pay under her own beliefs. The disposition to do so anyway is curiosity directed at hypotheses she does not yet hold, or a commitment made before the evidence, as in a pre-analysis plan. Scientific practice pairs multiplicity corrections with pre-registration and replication for this reason, and these are the complements the result identifies, not a broader critical thinking.

Creative thinking can offset the discount, but only in particular forms. The discount grows with the number of candidates searched, so generating more candidates after seeing the evidence raises it. What lowers the surprise a correct template needs is a narrower search, a template held before the evidence, or a template whose parameter is common to many situations, so that the evidence from all of them bears on it at once.

Corollary 5 (Creative thinking and the discount)**.** *Under the conditions of Proposition 13, let x contain observations from N situations that are independent given each node and share the node's predictive covariance V_S, and let t be a centered invariant template whose parameter is common to the N situations, with precision υ in each. Let $z_{t,\ell}$ be the standardized surprise along t in situation ℓ, and $Z_t := N^{-1/2}\sum_\ell z_{t,\ell}$. Then:*

(i) The critical thinker admits t if and only if $Z_t^2 > c_\kappa(N\upsilon, |\mathscr{S}|)$.

(ii) The additional surprise critical thinking requires, $c_\kappa(N\upsilon,|\mathscr{S}|) - c_0(N\upsilon,|\mathscr{S}|) = 2\kappa\frac{1+N\upsilon}{N\upsilon}\log|\mathscr{S}|$, is zero when $|\mathscr{S}| = 1$, increasing in $|\mathscr{S}|$, and decreasing in N.

(iii) If t is correct, with parameter $\vartheta_t \neq 0$, the probability that the critical thinker admits it tends to one as $N \to \infty$, for every $\kappa \in [0,1]$ and every $|\mathscr{S}|$.

Part (ii) makes the three routes precise: the discount vanishes for a template held before the evidence, falls with a narrower search, and falls as a template applies to more situations. Part (iii) shows that an invariant template applied widely enough is eventually admitted however strong the discount. Panel B of Figure 8 shows the cost and its remedy: at small N the critical thinker admits a correct template far less often than the naive DM, and pooling closes the gap. Pooling has a second effect: by Proposition 13 with precision $N\upsilon$, the threshold value of κ that keeps false insights out rises to $N\upsilon/(1+N\upsilon)$. A full discount, $\kappa = 1$, still meets it,

so a DM who discounts fully and applies templates widely both keeps chance fits off the menu and admits the correct template. Critical thinking thus raises the return to creative thinking that narrows the search or pools evidence across situations, and lowers the return to creative thinking that only adds candidates. Compression and pruning are complements: pruning makes compression worth more, and compression is what lets a correct insight survive pruning.

7 Concluding Remarks

This paper asks what stimulates curiosity: the drive to seek the information required to explore new perspectives. It treats contextual shifts in mindsets as a choice of how much model uncertainty to entertain. After an initial signal, a DM with a familiar default and latent, less precise alternatives jointly chooses how much weight to place on each and what information to acquire, disciplined by an evidence constraint that caps the weight on any alternative at its Bayesian posterior. Curiosity arises endogenously as the value of resolving the model uncertainty she entertains. Because she can seek information, exploration is both triggered and tested by evidence, which helps characterize conditions under which hallmark effects of associative-memory models persist. Payoff-irrelevant cues that leave the default unchanged sway decisions only when they activate an alternative under which they are relevant, and the shift lasts only if evidence supports it; yet rival interpretations can stifle exploration of a promising alternative by dividing the evidence, even if she would never entertain them. Endogenizing the menu, creative thinking compresses the space of mental models into reusable templates and critical thinking prunes it to those that survive evidence not used to form them. Two results depart from common sense: how far a mental model transfers is fixed by its templates rather than chosen, so a template that explains why pays only through transfer, which it does not by itself deliver; and the pitfall of critical thinking is a curiosity failure rather than a rejection.

Taken together, the results suggest that open-mindedness, curiosity, creative thinking, and critical thinking form a capability growth cycle within the brain's evidence-based reasoning (Figure 9). Open-mindedness, entertaining an alternative to the extent the evidence allows, is what curiosity requires, since there is no model uncertainty to resolve until an alternative is entertained. Curiosity tends to extinguish itself, since the evidence it gathers resolves the uncertainty that prompted it. Creative thinking can replenish that uncertainty: the observations curiosity seeks were not used to form the mental models, so their surprises can justify a new one, and they are the fresh tests critical thinking requires. Critical thinking keeps chance fits off the menu, where they would dilute the evidence cap of the right alternative, but it can leave a correct template untested unless curiosity, or evidence that arrives unsought, supplies fresh evidence. And a mental model that sets the cycle going in one situation does not carry over for free: extending it is a horizontal step that the evidence must justify and that costs mental effort. These results lay microfoundations for capability formation rather than a full theory of how capabilities accumulate over a lifetime; deriving that dynamic process is left for future research.

The cycle matters most where the situations people face keep changing, as in training developing-country teachers in new pedagogies that go beyond transmission of subject content, such as inquiry-based instruction. These pedagogies often ask teachers to connect knowledge to practical needs that differ across classrooms and shift as developing economies transform. Since teaching why a practice works does not by itself make trainees generalize, the framework points to an integrative approach with two parts. The first is to target templates. Training cannot place a mental model on a trainee's menu, since only she constructs it; it can activate one she already holds and shape the templates she builds new ones from. This is the distinguishing feature of the "Learning to Learn" approach to teacher training (Nourani et al., 2025, 2026), which works through guiding texts organized around capabilities rather than lesson plans, presents content as models with a limited range of validity, and carries concepts deliberately across domains. Reflection on inquiry-based against traditional instruction sharpens these templates; simulation and discussion with peers let teachers see the mental models they assemble work until they are confident enough to explore their consequences in their own classrooms;

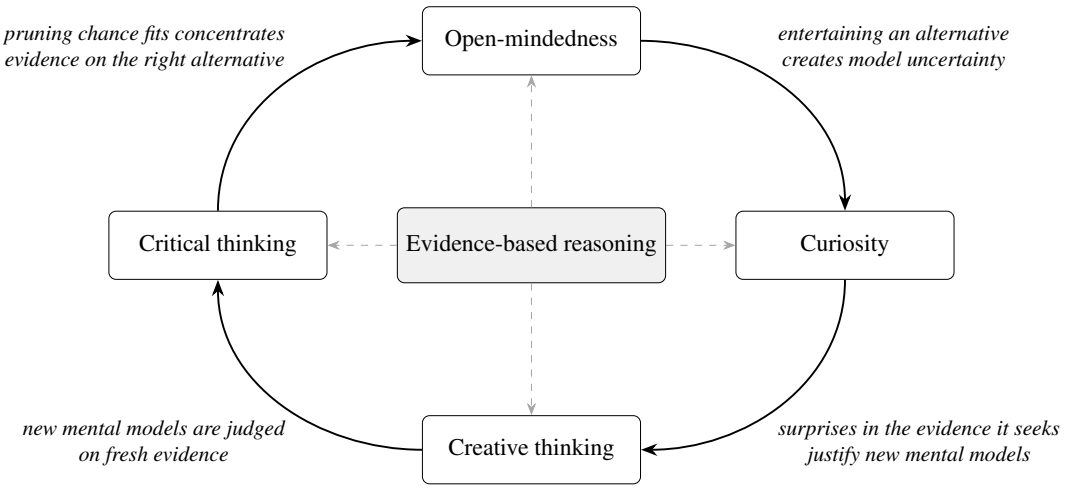


FIGURE 9. THE CAPABILITY GROWTH CYCLE

Notes: Each solid arrow summarizes a result; the dashed arrows indicate that every process is disciplined by the evidence constraint. Curiosity extinguishes itself as the evidence it gathers resolves model uncertainty, so the cycle continues only if creative thinking replenishes it.

drawing analogies trains the horizontal steps; and separating facts from opinions is the discipline against after-the-fact fitting formalized here. The second part is a stream of opportunities that keeps the capability growth cycle going. Because a shift to a new mental model must promise enough to pay for itself, opportunities must raise the returns to the new pedagogy and make evidence on them accessible. School-based agricultural education supplies both as a platform of experiential science learning and high-return farming techniques (Lee, 2026): school farms serve as demonstration plots, science lessons and student home projects make evidence cheap and draw it from many plots at once, and each new challenge, such as a pest outbreak or a new crop, calls for a new mental model that the stream of evidence lets teachers build and test. Mental models with lower but more certain payoffs then serve as stepping stones to ones with higher but less certain payoffs, and capabilities keep growing. My ongoing research, which restructures school-based agricultural education around the Learning to Learn approach in a randomized trial across Liberian schools, is designed to test the value of this integration: how much the stream of opportunities adds to the training.

References

Abdellaoui, Mohammed, Aurélien Baillon, Laetitia Placido, and Peter P. Wakker. 2011. "The Rich Domain of Uncertainty: Source Functions and Their Experimental Implementation." *American Economic Review* 101 (2): 695–723.

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In *Handbook of Labor Economics*, Vol. 4B, edited by David Card and Orley Ashenfelter, 1043–171. Amsterdam: Elsevier.

Aghion, Philippe, and Peter Howitt. 1992. "A Model of Growth through Creative Destruction." *Econometrica* 60 (2): 323–51.

Almlund, Mathilde, Angela Lee Duckworth, James Heckman, and Tim Kautz. 2011. "Personality Psychology and Economics." In *Handbook of the Economics of Education*, Vol. 4, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 1–181. Amsterdam: Elsevier.

Alsan, Marcella, and Marianne Wanamaker. 2018. "Tuskegee and the Health of Black Men." *Quarterly Journal of Economics* 133 (1): 407–55.

Anderson, Michael C. 2003. "Rethinking Interference Theory: Executive Control and the Mechanisms of Forgetting." *Journal of Memory and Language* 49 (4): 415–45.

Anderson, Michael C., and James H. Neely. 1996. "Interference and Inhibition in Memory Retrieval." In *Memory*, edited by Elizabeth L. Bjork and Robert A. Bjork, 237–313. San Diego: Academic Press.

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103 (484): 1481–95.

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart. 2022. "Subjective Models of the Macroeconomy: Evidence from Experts and Representative Samples." *Review of Economic Studies* 89 (6): 2958–91.

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics* 118 (4): 1279–333.

Ba, Cuimin. 2026. "Robust Misspecified Models." *American Economic Review* 116 (4): 1340–79.

Baillon, Aurélien, Zhenxing Huang, Asli Selim, and Peter P. Wakker. 2018. "Measuring Ambiguity Attitudes for All (Natural) Events." *Econometrica* 86 (5): 1839–58.

Banerjee, Abhijit, Swati Bhattacharjee, Raghabendra Chattopadhyay, Esther Duflo, Alejandro J. Ganimian, Kailash Rajah, and Elizabeth S. Spelke. 2025. "Children's Arithmetic Skills Do Not Transfer between Applied and Academic Mathematics." *Nature* 639 (8055): 673–81.

Banerjee, Abhijit V., and Andrew F. Newman. 1993. "Occupational Choice and the Process of Development." *Journal of Political Economy* 101 (2): 274–98.

Bays, Paul M., and Masud Husain. 2008. "Dynamic Shifts of Limited Working Memory Resources in Human Vision." *Science* 321 (5890): 851–54.

Becker, Gary S. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. New York: National Bureau of Economic Research.

Behrens, Timothy E. J., Timothy H. Muller, James C. R. Whittington, Shirley Mark, Alon B. Baram, Kimberly L. Stachenfeld, and Zeb Kurth-Nelson. 2018. "What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior." *Neuron* 100 (2): 490–509.

Bernardi, Silvia, Marcus K. Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C. Daniel Salzman. 2020. "The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex." *Cell* 183 (4): 954–67.

Bernheim, B. Douglas. 2009. "On the Potential of Neuroeconomics: A Critical (but Hopeful) Appraisal." *American Economic Journal: Microeconomics* 1 (2): 1–41.

Bold, Tessa, Kayuki C. Kaizzi, Jakob Svensson, and David Yanagizawa-Drott. 2017. "Lemon Technologies and Adoption: Measurement, Theory, and Evidence from Agricultural Markets in Uganda." *Quarterly Journal of Economics* 132 (3): 1055–1100.

Bénabou, Roland, and Jean Tirole. 2016. "Mindful Economics: The Production, Consumption, and Value of Beliefs." *Journal of Economic Perspectives* 30 (3): 141–64.

Blackwell, David. 1953. "Equivalent Comparisons of Experiments." *Annals of Mathematical Statistics* 24 (2): 265–72.

Boden, Margaret A. 2004. *The Creative Mind: Myths and Mechanisms*. 2nd ed. London: Routledge.

Bordalo, Pedro, John J. Conlon, Nicola Gennaioli, Spencer Y. Kwon, and Andrei Shleifer. 2023. "Memory and Probability." *Quarterly Journal of Economics* 138 (1): 265–311.

Bordalo, Pedro, Giovanni Burro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2025. "Imagining the Future: Memory, Simulation, and Beliefs." *Review of Economic Studies* 92 (3): 1532–63.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2020. "Memory, Attention, and Choice." *Quarterly Journal of Economics* 135 (3): 1399–1442.

Bordalo, Pedro, Nicola Gennaioli, Giacomo Lanzani, and Andrei Shleifer. 2026. "A Cognitive Theory of Reasoning and Choice." *Quarterly Journal of Economics* 141 (3): 1921–63.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas ter Weel. 2008. "The Economics and Psychology of Personality Traits." *Journal of Human Resources* 43 (4): 972–1059.

Brock, William A., Steven N. Durlauf, and Kenneth D. West. 2003. "Policy Evaluation in Uncertain Economic Environments." *Brookings Papers on Economic Activity* 2003 (1): 235–322.

Brunnermeier, Markus K., and Jonathan A. Parker. 2005. "Optimal Expectations." *American Economic Review* 95 (4): 1092–1118.

Campbell, Donald T. 1960. "Blind Variation and Selective Retention in Creative Thought as in Other Knowledge Processes." *Psychological Review* 67 (6): 380–400.

Caplin, Andrew, and Mark Dean. 2015. "Revealed Preference, Rational Inattention, and Costly Information Acquisition." *American Economic Review* 105 (7): 2183–2203.

Cogley, Timothy, and Thomas J. Sargent. 2005. "The Conquest of US Inflation: Learning and Robustness to Model Uncertainty." *Review of Economic Dynamics* 8 (2): 528–63.

Colgin, Laura Lee, Edvard I. Moser, and May-Britt Moser. 2008. "Understanding Memory through Hippocampal Remapping." *Trends in Neurosciences* 31 (9): 469–77.

Constantinescu, Alexandra O., Jill X. O'Reilly, and Timothy E. J. Behrens. 2016. "Organizing Conceptual Knowledge in Humans with a Gridlike Code." *Science* 352 (6292): 1464–68.

Cowan, Nelson. 2001. "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity." *Behavioral and Brain Sciences* 24 (1): 87–114.

Cunha, Flavio, and James J. Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97 (2): 31–47.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931.

Davies, Benjamin, and Anirudh Sankar. 2026. "The Value of Conceptual Knowledge." arXiv:2509.09170.

Dragoi, George, and Susumu Tonegawa. 2011. "Preplay of Future Place Cell Sequences by Hippocampal Cellular Assemblies." *Nature* 469 (7330): 397–401.

Dubey, Rachit, and Thomas L. Griffiths. 2020. "Reconciling Novelty and Complexity through a Rational Analysis of Curiosity." *Psychological Review* 127 (3): 455–76.

Eliaz, Kfir, and Ran Spiegler. 2020. "A Model of Competing Narratives." *American Economic Review* 110 (12): 3786–3816.

Enke, Benjamin, Frederik Schwerter, and Florian Zimmermann. 2024. "Associative Memory, Beliefs and Market Interactions." *Journal of Financial Economics* 157.

Ennis, Robert H. 1987. "A Taxonomy of Critical Thinking Dispositions and Abilities." In *Teaching Thinking Skills: Theory and Practice*, edited by Joan B. Baron and Robert J. Sternberg, 9–26. New York: W. H. Freeman.

Esponda, Ignacio, and Demian Pouzo. 2016. "Berk–Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models." *Econometrica* 84 (3): 1093–1130.

Foster, Andrew D., and Mark R. Rosenzweig. 1996. "Technical Change and Human-Capital Returns and Investments: Evidence from the Green Revolution." *American Economic Review* 86 (4): 931–53.

Galor, Oded, and Joseph Zeira. 1993. "Income Distribution and Macroeconomics." *Review of Economic Studies* 60 (1): 35–52.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102 (6): 460–65.

Gennaioli, Nicola, and Andrei Shleifer. 2010. "What Comes to Mind." *Quarterly Journal of Economics* 125 (4): 1399–1433.

Gittins, John C. 1979. "Bandit Processes and Dynamic Allocation Indices." *Journal of the Royal Statistical Society, Series B* 41 (2): 148–77.

Godden, Duncan R., and Alan D. Baddeley. 1975. "Context-Dependent Memory in Two Natural Environments: On Land and Underwater." *British Journal of Psychology* 66 (3): 325–31.

Golman, Russell, and George Loewenstein. 2018. "Information Gaps: A Theory of Preferences Regarding the Presence and Absence of Information." *Decision* 5 (3): 143–64.

Golman, Russell, David Hagmann, and George Loewenstein. 2017. "Information Avoidance." *Journal of Economic Literature* 55 (1): 96–135.

Griffiths, Thomas L., and Joshua B. Tenenbaum. 2009. "Theory-Based Causal Induction." *Psychological Review* 116 (4): 661–716.

Gul, Faruk, and Wolfgang Pesendorfer. 2008. "The Case for Mindless Economics." In *The Foundations of Positive and Normative Economics: A Handbook*, edited by Andrew Caplin and Andrew Schotter, 3–42. Oxford: Oxford

University Press.

Gupta, Anoopum S., Matthijs A. A. van der Meer, David S. Touretzky, and A. David Redish. 2010. "Hippocampal Replay Is Not a Simple Function of Experience." *Neuron* 65 (5): 695–705.

Hafting, Torkel, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I. Moser. 2005. "Microstructure of a Spatial Map in the Entorhinal Cortex." *Nature* 436 (7052): 801–6.

Halpern, Diane F. 1998. "Teaching Critical Thinking for Transfer across Domains: Disposition, Skills, Structure Training, and Metacognitive Monitoring." *American Psychologist* 53 (4): 449–55.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. 2014. "Learning Through Noticing: Theory and Evidence from a Field Experiment." *Quarterly Journal of Economics* 129 (3): 1311–53.

Hansen, Lars Peter, and Thomas J. Sargent. 2008. *Robustness*. Princeton, NJ: Princeton University Press.

Hassabis, Demis, Dharshan Kumaran, Seralynne D. Vann, and Eleanor A. Maguire. 2007. "Patients with Hippocampal Amnesia Cannot Imagine New Experiences." *Proceedings of the National Academy of Sciences* 104 (5): 1726–31.

Heckman, James J., and Chase O. Corbin. 2016. "Capabilities and Skills." *Journal of Human Development and Capabilities* 17 (3): 342–59.

Heckman, James J., and Tim Kautz. 2012. "Hard Evidence on Soft Skills." *Labour Economics* 19 (4): 451–64.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24 (3): 411–82.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack. 2018. "Unrealistic Expectations and Misguided Learning." *Econometrica* 86 (4): 1159–1214.

Helpman, Elhanan, and Manuel Trajtenberg. 1998. "A Time to Sow and a Time to Reap: Growth Based on General Purpose Technologies." In *General Purpose Technologies and Economic Growth*, edited by Elhanan Helpman, 55–83. Cambridge, MA: MIT Press.

Howard, Marc W., and Michael J. Kahana. 2002. "A Distributed Representation of Temporal Context." *Journal of Mathematical Psychology* 46 (3): 269–99.

Howitt, Peter. 1999. "Steady Endogenous Growth with Population and R&D Inputs Growing." *Journal of Political Economy* 107 (4): 715–30.

Jefferys, William H., and James O. Berger. 1992. "Ockham's Razor and Bayesian Analysis." *American Scientist* 80 (1): 64–72.

Jezek, Karel, Espen J. Henriksen, Alessandro Treves, Edvard I. Moser, and May-Britt Moser. 2011. "Theta-Paced Flickering between Place-Cell Maps in the Hippocampus." *Nature* 478 (7368): 246–49.

Kahana, Michael J. 2012. *Foundations of Human Memory*. New York: Oxford University Press.

Kang, Min Jeong, Ming Hsu, Ian M. Krajbich, George Loewenstein, Samuel M. McClure, Joseph Tao-yi Wang, and Colin F. Camerer. 2009. "The Wick in the Candle of Learning: Epistemic Curiosity Activates Reward Circuitry and Enhances Memory." *Psychological Science* 20 (8): 963–73.

Karni, Edi, and Marie-Louise Vierø. 2013. "Reverse Bayesianism: A Choice-Based Theory of Growing Awareness." *American Economic Review* 103 (7): 2790–2810.

Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90 (430): 773–95.

Kemp, Charles, and Joshua B. Tenenbaum. 2008. "The Discovery of Structural Form." *Proceedings of the National Academy of Sciences* 105 (31): 10687–92.

Khaw, Mel Win, Ziang Li, and Michael Woodford. 2021. "Cognitive Imprecision and Small-Stakes Risk Aversion." *Review of Economic Studies* 88 (4): 1979–2013.

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji. 2005. "A Smooth Model of Decision Making under Ambiguity." *Econometrica* 73 (6): 1849–92.

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji. 2009. "Recursive Smooth Ambiguity Preferences." *Journal of Economic Theory* 144 (3): 930–76.

Knoblich, Günther, Stellan Ohlsson, Hilde Haider, and Detlef Rhenius. 1999. "Constraint Relaxation and Chunk Decomposition in Insight Problem Solving." *Journal of Experimental Psychology: Learning, Memory, and Cognition*

25 (6): 1534–55.

Kobayashi, Kenji, Silvio Ravaioli, Adrien Baranès, Michael Woodford, and Jacqueline Gottlieb. 2019. "Diverse Motives for Human Curiosity." *Nature Human Behaviour* 3 (6): 587–95.

Koestler, Arthur. 1964. *The Act of Creation*. London: Hutchinson.

Kool, Wouter, Joseph T. McGuire, Zev B. Rosen, and Matthew M. Botvinick. 2010. "Decision Making and the Avoidance of Cognitive Demand." *Journal of Experimental Psychology: General* 139 (4): 665–82.

Kreps, David M., and Evan L. Porteus. 1978. "Temporal Resolution of Uncertainty and Dynamic Choice Theory." *Econometrica* 46 (1): 185–200.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108 (3): 480–98.

Laajaj, Rachid, and Karen Macours. 2026. "The Complexity of Multidimensional Learning in Agriculture." *Econometrica* 94 (2): 465–503.

Lee, Jimmy. 2026. "Misinterpretation as an Information Friction? Activating School-Based Agricultural Education in Liberia with Parental Engagement." Job market paper, Northwestern University.

Leopold, David A., and Nikos K. Logothetis. 1999. "Multistable Phenomena: Changing Views in Perception." *Trends in Cognitive Sciences* 3 (7): 254–63.

Leutgeb, Jill K., Stefan Leutgeb, Alessandro Treves, Retsina Meyer, Carol A. Barnes, Bruce L. McNaughton, May-Britt Moser, and Edvard I. Moser. 2005. "Progressive Transformation of Hippocampal Neuronal Representations in 'Morphed' Environments." *Neuron* 48 (2): 345–58.

Li, Jian. 2020. "Preferences for Partial Information and Ambiguity." *Theoretical Economics* 15 (3): 1059–94.

Lieder, Falk, and Thomas L. Griffiths. 2020. "Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources." *Behavioral and Brain Sciences* 43: e1.

Liquin, Emily G., and Tania Lombrozo. 2020. "A Functional Approach to Explanation-Seeking Curiosity." *Cognitive Psychology* 119: 101276.

List, John A., Azeem M. Shaikh, and Yang Xu. 2019. "Multiple Hypothesis Testing in Experimental Economics." *Experimental Economics* 22 (4): 773–93.

Loewenstein, George. 1994. "The Psychology of Curiosity: A Review and Reinterpretation." *Psychological Bulletin* 116 (1): 75–98.

MacKay, David J. C. 1992. "Bayesian Interpolation." *Neural Computation* 4 (3): 415–47.

Maćkowiak, Bartosz, Filip Matějka, and Mirko Wiederholt. 2023. "Rational Inattention: A Review." *Journal of Economic Literature* 61 (1): 226–73.

Marinacci, Massimo. 2015. "Model Uncertainty." *Journal of the European Economic Association* 13 (6): 1022–1100.

Matějka, Filip, and Alisdair McKay. 2015. "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model." *American Economic Review* 105 (1): 272–98.

Michelson, Hope, Anna Fairbairn, Brenna Ellison, Annemie Maertens, and Victor Manyong. 2021. "Misperceived Quality: Fertilizer in Tanzania." *Journal of Development Economics* 148: 102579.

Mookherjee, Dilip, and Debraj Ray. 2003. "Persistent Inequality." *Review of Economic Studies* 70 (2): 369–93.

Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer. 2008. "Coarse Thinking and Persuasion." *Quarterly Journal of Economics* 123 (2): 577–619.

Munshi, Kaivan. 2004. "Social Learning in a Heterogeneous Population: Technology Diffusion in the Indian Green Revolution." *Journal of Development Economics* 73 (1): 185–213.

Nelson, Richard R., and Edmund S. Phelps. 1966. "Investment in Humans, Technological Diffusion, and Economic Growth." *American Economic Review* 56 (1/2): 69–75.

Nickerson, Raymond S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2 (2): 175–220.

Nourani, Vesall, Nava Ashraf, and Abhijit Banerjee. 2025. "Learning to Teach by Learning to Learn." Working paper, Harvard University.

Nourani, Vesall, Erin Murphy-Graham, Joseph Lample, and Sarah Kabay. 2026. "Transformative Teacher Professional Development: A Qualitative Case Study of a Metacognitive Approach to Teacher Training in Uganda." *Teaching*

and Teacher Education.

O'Keefe, John, and Jonathan Dostrovsky. 1971. "The Hippocampus as a Spatial Map: Preliminary Evidence from Unit Activity in the Freely-Moving Rat." *Brain Research* 34 (1): 171–75.

O'Keefe, John, and Lynn Nadel. 1978. *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press.

Olken, Benjamin A. 2015. "Promises and Perils of Pre-Analysis Plans." *Journal of Economic Perspectives* 29 (3): 61–80.

Ortoleva, Pietro. 2012. "Modeling the Change of Paradigm: Non-Bayesian Reactions to Unexpected News." *American Economic Review* 102 (6): 2410–36.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press.

Pearl, Judea, and Elias Bareinboim. 2014. "External Validity: From Do-Calculus to Transportability across Populations." *Statistical Science* 29 (4): 579–95.

Peters, Jonas, Peter Bühlmann, and Nicolai Meinshausen. 2016. "Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals." *Journal of the Royal Statistical Society, Series B* 78 (5): 947–1012.

Pratt, John W. 1964. "Risk Aversion in the Small and in the Large." *Econometrica* 32 (1–2): 122–36.

Redlawsk, David P., Andrew J. W. Civettini, and Karen M. Emmerson. 2010. "The Affective Tipping Point: Do Motivated Reasoners Ever 'Get It'?" *Political Psychology* 31 (4): 563–93.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5): S71–S102.

Sala, Giovanni, and Fernand Gobet. 2017. "Does Far Transfer Exist? Negative Evidence from Chess, Music, and Working Memory Training." *Current Directions in Psychological Science* 26 (6): 515–20.

Sanders, Honi, Matthew A. Wilson, and Samuel J. Gershman. 2020. "Hippocampal Remapping as Hidden State Inference." *eLife* 9: e51140.

Sankar, Anirudh, Robert Dulin, Benjamin Davies, Vesall Nourani, Jess Rudder, Abraham Salomon, and Godfrey Taulya. 2026. "How Mechanistic Explanations Reshape Learning and Behavior: Evidence from a Fertilizer Choice Experiment in Eastern Uganda." Working paper, Stanford University.

Schacter, Daniel L., and Donna Rose Addis. 2007. "The Cognitive Neuroscience of Constructive Memory: Remembering the Past and Imagining the Future." *Philosophical Transactions of the Royal Society B* 362 (1481): 773–86.

Schultz, Theodore W. 1975. "The Value of the Ability to Deal with Disequilibria." *Journal of Economic Literature* 13 (3): 827–46.

Schwartzstein, Joshua. 2014. "Selective Attention and Learning." *Journal of the European Economic Association* 12 (6): 1423–52.

Schwartzstein, Joshua, and Adi Sunderam. 2021. "Using Models to Persuade." *American Economic Review* 111 (1): 276–323.

Shenhav, Amitai, Matthew M. Botvinick, and Jonathan D. Cohen. 2013. "The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function." *Neuron* 79 (2): 217–40.

Shenhav, Amitai, Sebastian Musslick, Falk Lieder, Wouter Kool, Thomas L. Griffiths, Jonathan D. Cohen, and Matthew M. Botvinick. 2017. "Toward a Rational and Mechanistic Account of Mental Effort." *Annual Review of Neuroscience* 40: 99–124.

Shiller, Robert J. 2017. "Narrative Economics." *American Economic Review* 107 (4): 967–1004.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66.

Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50 (3): 665–90.

Skinner, B. F. 1948. "'Superstition' in the Pigeon." *Journal of Experimental Psychology* 38 (2): 168–72.

Smith, Steven M., and Edvard Vela. 2001. "Environmental Context-Dependent Memory: A Review and Meta-Analysis." *Psychonomic Bulletin and Review* 8 (2): 203–20.

Spelke, Elizabeth S., and Katherine D. Kinzler. 2007. "Core Knowledge." *Developmental Science* 10 (1): 89–96.

Spiegler, Ran. 2016. "Bayesian Networks and Boundedly Rational Expectations." *Quarterly Journal of Economics* 131 (3): 1243–90.

Stanovich, Keith E., and Richard F. West. 1997. "Reasoning Independently of Prior Belief and Individual Differences in

Actively Open-Minded Thinking." *Journal of Educational Psychology* 89 (2): 342–57.

Stantcheva, Stefanie. 2021. "Understanding Tax Policy: How Do People Reason?" *Quarterly Journal of Economics* 136 (4): 2309–69.

Strzalecki, Tomasz. 2013. "Temporal Resolution of Uncertainty and Recursive Models of Ambiguity Aversion." *Econometrica* 81 (3): 1039–74.

Tenenbaum, Joshua B., Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. "How to Grow a Mind: Statistics, Structure, and Abstraction." *Science* 331 (6022): 1279–85.

Tolman, Edward C. 1948. "Cognitive Maps in Rats and Men." *Psychological Review* 55 (4): 189–208.

Tulving, Endel, and Donald M. Thomson. 1973. "Encoding Specificity and Retrieval Processes in Episodic Memory." *Psychological Review* 80 (5): 352–73.

Tversky, Amos, and Derek J. Koehler. 1994. "Support Theory: A Nonextensional Representation of Subjective Probability." *Psychological Review* 101 (4): 547–67.

Underwood, Benton J. 1957. "Interference and Forgetting." *Psychological Review* 64 (1): 49–60.

Webb, Ryan, Paul W. Glimcher, and Kenway Louie. 2021. "The Normalization of Consumer Valuations: Context-Dependent Preferences from Neurobiological Constraints." *Management Science* 67 (1): 93–125.

Weitzman, Martin L. 1979. "Optimal Search for the Best Alternative." *Econometrica* 47 (3): 641–54.

Weitzman, Martin L. 1998. "Recombinant Growth." *Quarterly Journal of Economics* 113 (2): 331–60.

Whittington, James C. R., Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E. J. Behrens. 2020. "The Tolman–Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation." *Cell* 183 (5): 1249–63.

Willingham, Daniel T. 2007. "Critical Thinking: Why Is It So Hard to Teach?" *American Educator* 31 (2): 8–19.

Wills, Tom J., Colin Lever, Francesca Cacucci, Neil Burgess, and John O'Keefe. 2005. "Attractor Dynamics in the Hippocampal Representation of the Local Environment." *Science* 308 (5723): 873–76.

Woodford, Michael. 2012. "Prospect Theory as Efficient Perceptual Distortion." *American Economic Review* 102 (3): 41–46.

Zhang, Huanqiu, P. Dylan Rich, Albert K. Lee, and Tatyana O. Sharpee. 2023. "Hippocampal Spatial Representations Exhibit a Hyperbolic Geometry That Expands with Experience." *Nature Neuroscience* 26 (1): 131–39.

Berk, Robert H. 1966. "Limiting Behavior of Posterior Distributions when the Model is Incorrect." *Annals of Mathematical Statistics* 37 (1): 51–58.

Topkis, Donald M. 1978. "Minimizing a Submodular Function on a Lattice." *Operations Research* 26 (2): 305–21.

A Proofs

Throughout, ϕ_ν denotes the $\mathcal{N}(0,\nu^2)$ density, $h_k := b_k'\Sigma^{-1}b_k = \nu_k^{-2}$ the precision with which model k's direction can be read, and "the two-model comparison" the evidence constraint (5) restricted to two nodes, under which a node's evidence cap exceeds one half if and only if its log Bayes factor against the other is positive.

Proofs follow the order of the text. Statements in the Online Appendix are proved where they are stated.

Mental models, signals, and timing (Section 2.2.1).

Proof of Lemma 1. Under (1) the log-likelihood of x given θ is, up to terms free of θ,

$$\theta\, b_k'\Sigma^{-1}x - \tfrac{1}{2}\theta^2 h_k \;=\; -\tfrac{h_k}{2}\,(s_k-\theta)^2 + \text{const}, \qquad s_k := \frac{b_k'\Sigma^{-1}x}{h_k} = \alpha_k' x.$$

By the factorization theorem s_k is sufficient for θ, and the display shows $s_k \mid \theta \sim \mathcal{N}(\theta,\nu_k^2)$. Conjugate normal updating from the prior $\mathcal{N}(\mu_k,\sigma_k^2)$ gives the stated posterior. □

Specifying the optimization problem in closed form (Section 2.3).

Proof of the closed forms (9)–(12). *The variance* $\psi_k(n)$. The posterior mean is a martingale under model k, so its ex-ante variance is the reduction in posterior variance, $1/P_k(0)-1/P_k(n)$, increasing and concave in n since P_k is affine in n. *Equation* (9). Given a step-(4) posterior mean M under model k, (8) gives $a^* = (M-\chi)/\lambda$ and value $(M-\chi)^2/(2\lambda)$. Taking the step-(2) expectation over $M \sim \mathcal{N}(m_k,\psi_k(n))$ yields

$$U_k = \frac{(m_k-\chi)^2+\psi_k(n)}{2\lambda}, \qquad \partial_n U_k = \frac{\psi_k'(n)}{2\lambda} = \frac{1}{2\lambda\,\nu_y^2 P_k(n)^2}.$$

Equations (10)–(12). With φ the identity the single-act problem $\max_a\{a(\bar m-\chi)-\frac{\lambda}{2}a^2\}$ has value $(\bar m-\chi)^2/(2\lambda)$, whereas the model-by-model term in (7) is $\sum_k w_k(m_k-\chi)^2/(2\lambda)$. The difference is $\mathrm{Var}_{\mathbf{w}}(m)/(2\lambda)$, which is the hedging cost stated in the text. The single-act term of (4) is therefore $\sum_k w_k U_k(x) - A(\mathbf{w})$, and adding the within-model term $\sum_k w_k[U_k(x,n)-U_k(x)]$ gives (11). With one alternative held, the two-point variance is $w_k(1-w_k)(m_k-m_0)^2$, giving (10); the cost (3) reduces to $(\gamma/2)w_k^2$; and (11) becomes (12). Both $D_k(x)$ and $\overline{W}_k(x)$ depend on x, the former through $m_k(s_k)$ (Lemma 1) and the latter through (5). □

Main Results (Section 3).

Proof of Lemma 2. Using $2\sum_{j<k}w_jw_k = W^2-\sum_k w_k^2$, write $C(\mathbf{w}) = \frac{\eta}{2}W^2 - \frac{\eta-\gamma}{2}\sum_{k\geq 1}w_k^2$. For fixed W, minimizing C is maximizing the strictly convex function $\sum_k w_k^2$ over the simplex $\{\mathbf{w}\geq 0 : \sum_k w_k = W\}$, which is attained only at a vertex, where $C=\frac{\gamma}{2}W^2$. Elsewhere $C(\mathbf{w})-\frac{\gamma}{2}W^2 = (\eta-\gamma)\sum_{j<k}w_jw_k > 0$. □

Alternative activation (Section 3.1).

Proof of Proposition 1. From (12),

$$\partial_{w_k}V = \Delta U_k - D_k(1-2w_k) - \gamma w_k, \qquad \partial^2_{w_kw_k}V = 2D_k-\gamma,$$

so the objective is quadratic in w_k with constant curvature, and the problem is to maximize it on $[0,\overline{W}_k]$.

Part (ii)(a). If $\gamma > 2D_k$, V is strictly concave. Its unconstrained maximizer is $(\Delta U_k - D_k)/(\gamma-2D_k)$, and projecting onto $[0,\overline{W}_k]$ gives the median expression, the unique maximizer on the interval. Since $\partial_{w_k}V|_{w_k=0} = \Delta U_k - D_k$, the optimum is zero exactly when $\Delta U_k \leq D_k$, and it is continuous and nondecreasing in ΔU_k.

Part (ii)(b). If $\gamma < 2D_k$, V is strictly convex, so the maximum is at an endpoint, and

$$V(\overline{W}_k) - V(0) \;=\; \overline{W}_k\big[\Delta U_k - D_k(1-\overline{W}_k) - \tfrac{\gamma}{2}\overline{W}_k\big],$$

which is positive if and only if $\Delta U_k > T_k(\overline{W}_k)$. Since $T_k'(\overline{W}_k) = \gamma/2 - D_k < 0$, the threshold falls with the cap, and $T_k(\overline{W}_k) < D_k$ for $\overline{W}_k > 0$. At the threshold the optimum jumps from 0 to $\overline{W}_k$. (At $\gamma = 2D_k$, V is linear and $w_k^* = \overline{W}_k$ if and only if $\Delta U_k > D_k$.)

Part (i). In each case $w_k^* = 0$ for $\Delta U_k \le T_k$ and $w_k^* > 0$ for $\Delta U_k > T_k$, with $T_k \le D_k$, and the optimum respects $w_k^* \le \overline{W}_k$ by construction. □

Complementarity and curiosity failures (Section 3.2).

Proof of Proposition 2. Differentiate (4) in w_k (with $w_0 = 1 - \sum_j w_j$) and then in n; only the within-model term depends on n, so

$$\partial^2_{w_k n} V \;=\; \varphi'(U_k)\,\partial_n U_k - \varphi'(U_0)\,\partial_n U_0,$$

which for linear φ is $\partial_n U_k - \partial_n U_0$. Substituting $\partial_n U_j = [2\lambda\nu_y^2 P_j(n)^2]^{-1}$ from (9) gives the displayed expression, positive because $P_k(n) < P_0(n)$ under the maintained assumption $P_k(0) < P_0(0)$. (For general φ the cross-partial is positive if and only if $(P_k(n)/P_0(n))^2 < \varphi'(U_k)/\varphi'(U_0)$.) A positive cross-partial makes V supermodular in (w_k, n), so $w_k^*(n)$ and $n^*(w_k)$ are nondecreasing (Topkis, 1978). Under Assumption 1, U_0 does not depend on n, so at $\mathbf{w} = 0$ signals have no value and $n^* = 0$. □

Proof of Corollary 1. With alternative k held alone under Assumption 1 and linear φ, the n-th signal raises the objective by $w_k[\psi_k(n) - \psi_k(n-1)]/(2\lambda) - c$, and $\psi_k(n) - \psi_k(n-1) = 1/P_k(n-1) - 1/P_k(n) = 1/(\nu_y^2 P_k(n-1)P_k(n))$, decreasing in n because P_k is increasing. Signals are bought up to the last one with positive gain, which gives the rule and its comparative statics. The continuous approximation solves $w_k\,\partial_n U_k = c$, that is, $P_k(n)^2 = w_k/(2\lambda\nu_y^2 c)$. □

The impact of the initial signal (Section 3.3).

Proof of Proposition 3. Write $\Omega_k := \Sigma + \sigma_k^2 b_k b_k'$ and $z_k := x - \mu_k b_k$. By the Sherman–Morrison formula and the matrix determinant lemma,

$$\Omega_k^{-1} = \Sigma^{-1} - \frac{\sigma_k^2\,\Sigma^{-1} b_k b_k' \Sigma^{-1}}{1 + \sigma_k^2 h_k}, \qquad \det\Omega_k = \det\Sigma\,(1 + \sigma_k^2 h_k).$$

Since $b_k'\Sigma^{-1} z_k = h_k(s_k - \mu_k)$, the quadratic form splits as $z_k'\Sigma^{-1} z_k = R_k + h_k(s_k - \mu_k)^2$, where $R_k = x'\Sigma^{-1}x - h_k s_k^2 \ge 0$ is the squared Σ^{-1}-norm of the component of x orthogonal to b_k. Hence $z_k'\Omega_k^{-1} z_k = R_k + q_k$ and $\log\det\Omega_k = \log\det\Sigma + c_k$, so $-2\log L_k(x) = d\log 2\pi + \log\det\Sigma + c_k + R_k + q_k$. Subtracting the expressions for k and 0 gives (13).

(i) $R_0 \equiv R_k$ if and only if the rank-one matrices $\Sigma^{-1}b_0 b_0'\Sigma^{-1}/h_0$ and $\Sigma^{-1}b_k b_k'\Sigma^{-1}/h_k$ coincide, that is, $b_0 \propto b_k$. (ii) Only $q_k = (s_k - \mu_k)^2/(\sigma_k^2 + \nu_k^2)$ involves μ_k, and it is the squared standardized distance of s_k from μ_k. (iii) c_k is strictly increasing in σ_k^2/ν_k^2. Finally, $\overline{W}_k = \Lambda_k/(1 + \sum_j \Lambda_j)$ is increasing in $\log\Lambda_k$, which is increasing in each of k's channel differences, and decreasing in $\log\Lambda_j$ for $j \ne k$; (5) is Bayes' rule under a uniform prior over the $K+1$ models. □

A formal definition of curiosity (Section 3.4).

Proof of Proposition 4. The decomposition. With φ linear, the act chosen at step (4) under the updated mixture is $a = (\bar{m}_n - \chi)/\lambda$, with payoff $(\bar{m}_n - \chi)^2/(2\lambda)$; without signals the payoff is $(\bar{m} - \chi)^2/(2\lambda)$. Under the mixture predictive, $\bar{m}_n$ is the posterior mean of θ and so a martingale, whence $\mathbb{E}_{\mathbf{w}}[(\bar{m}_n - \chi)^2] =$

$(\bar{m}-\chi)^2+\mathrm{Var}_{\mathbf{w}}(\bar{m}_n)$, the first equality in (14). For the second, the law of total variance gives $\mathrm{Var}_{\mathbf{w}}(\theta)=\mathbb{E}_{\mathbf{w}}[\mathrm{Var}(\theta\mid y^n)]+\mathrm{Var}_{\mathbf{w}}(\bar{m}_n)$, with

$$\mathrm{Var}_{\mathbf{w}}(\theta)=\sum_k \frac{w_k}{P_k(0)}+\mathrm{Var}_{\mathbf{w}}(m),\qquad \mathrm{Var}(\theta\mid y^n)=\sum_k \frac{w_{k,n}}{P_k(n)}+\mathrm{Var}_{\mathbf{w}_n}(m_{\cdot,n}).$$

Taking expectations, Lemma OA.2(ii) turns the first term of the conditional variance into $\sum_k w_k/P_k(n)$, and the second is $2\lambda\mathbb{E}_{\mathbf{w}}[A_n]$ by definition. Subtracting and dividing by 2λ yields (14) with $\psi_k(n)=1/P_k(0)-1/P_k(n)$.

Part (b)(i): shared signals. Let g_j be model j's step-(2) posterior density of θ and let θ_0 generate the signals. The predictive probability of y^n under model j is $L_j(y^n)=\int\prod_{i=1}^n \phi_{\nu_y}(y_i-\theta)\,g_j(\theta)\,d\theta$. Because the likelihood is Gaussian in θ, the Laplace approximation around the sample mean $\bar{y}_n$ is exact up to a factor $1+o(1)$:

$$L_j(y^n)\;=\;\prod_{i=1}^n \phi_{\nu_y}(y_i-\bar{y}_n)\;g_j(\bar{y}_n)\,\sqrt{2\pi\nu_y^2/n}\,(1+o(1))\quad\text{almost surely,}$$

uniformly in j, since g_0 and g_k are continuous and positive. The first and third factors are common to both models, so by the strong law ($\bar{y}_n\to\theta_0$) the Bayes factor converges, $\Lambda_k^y(y^n)\to g_k(\theta_0)/g_0(\theta_0)=:\Lambda_k^\infty\in(0,\infty)$, and (6) gives $w_{k,n}\to w_k\Lambda_k^\infty/(1-w_k+w_k\Lambda_k^\infty)$, an interior point. Posterior consistency within each model gives $m_{j,n}\to\theta_0$ for $j\in\{0,k\}$, hence $m_{k,n}-m_{0,n}\to 0$.

Part (b)(ii): alternative-specific signals. Under Assumption 1 the default assigns the signals a fixed density f_0 (with finite moments as needed), so $\log\Lambda_k^y(y^n)=\log L_k(y^n)-\sum_{i=1}^n\log f_0(y_i)$, and the expansion in part (b)(i), which holds along any sequence with $\bar{y}_n$ convergent, gives $\log L_k(y^n)=\sum_{i=1}^n\log\phi_{\nu_y}(y_i-\bar{y}_n)+\log g_k(\bar{y}_n)+\frac{1}{2}\log(2\pi\nu_y^2/n)+o(1)$. If the signals are drawn from $y_i=\theta_0+\varepsilon_i$, with density $f:=\phi_{\nu_y}(\cdot-\theta_0)$, the strong law gives $\bar{y}_n\to\theta_0$, $n^{-1}\sum_i\log\phi_{\nu_y}(y_i-\bar{y}_n)\to\mathbb{E}_f\log f$, and $n^{-1}\sum_i\log f_0(y_i)\to\mathbb{E}_f\log f_0$, while the remaining terms are $o(n)$. Hence $n^{-1}\log\Lambda_k^y\to\mathrm{KL}(f\,\|\,f_0)$, which is positive because $f_0\neq f$, and $w_{k,n}\to 1$. If instead the signals are generated by f_0, the same steps give $n^{-1}\log\Lambda_k^y\to\mathbb{E}_{f_0}[\log\phi_{\nu_y}(y-\theta^*)-\log f_0(y)]=-\mathrm{KL}(f_0\,\|\,\phi_{\nu_y}(\cdot-\theta^*))$, where $\theta^*=\mathbb{E}_{f_0}[y]$ is the pseudo-true value, the unique minimizer of $\mathrm{KL}(f_0\,\|\,\phi_{\nu_y}(\cdot-\theta))$ over the Gaussian location family (Berk, 1966). This is negative because f_0 lies outside the family, so $w_{k,n}\to 0$.

Part (a). If $w_k=1$ for some k, then $\mathbf{w}_n\equiv\mathbf{w}$ by (6) and $A=A_n=0$. For the limit, $A_n\to 0$ almost surely by part (b): with shared signals the posterior means merge, and with alternative-specific signals $w_{k,n}(1-w_{k,n})\to 0$. Since A_n is bounded by a multiple of $\max_{j,k}(m_{j,n}-m_{k,n})^2$ and each $m_{k,n}$ is a convex combination of m_k and the sample mean, whose fourth moment is bounded in n under Gaussian signals, $\sup_n\mathbb{E}_{\mathbf{w}}[A_n^2]<\infty$, so $\{A_n\}$ is uniformly integrable and $\mathbb{E}_{\mathbf{w}}[A_n]\to 0$ and the resolution component converges to $A(\mathbf{w})$. □

Ambiguity aversion (Section 3.5).

Proof of Proposition 5. (i) Write $S_\varphi(w):=\max_a[(1-w)\varphi(\mathbb{E}_0u(a,\theta))+w\varphi(\mathbb{E}_ku(a,\theta))]$, so that at $n=0$ the gain from holding k at weight w is $S_\varphi(w)-\varphi(U_0)-\frac{\gamma}{2}w^2$; under linear φ it is $S(w)-U_0-\frac{\gamma}{2}w^2$ with $S(w)=U_0+w\Delta U_k-w(1-w)D_k$. For $\tilde{\varphi}=h\circ\varphi$, h increasing and concave, $h'(\varphi(U_0))=1$, Jensen's inequality and concavity of h give

$$S_{\tilde{\varphi}}(w)\;\le\;\max_a h\big((1-w)\varphi(\mathbb{E}_0u)+w\varphi(\mathbb{E}_ku)\big)\;=\;h\big(S_\varphi(w)\big)\;\le\;\tilde{\varphi}(U_0)+S_\varphi(w)-\varphi(U_0),$$

so the gain under $\tilde{\varphi}$ is at most the gain under φ at every w; taking φ linear and $h=\tilde{\varphi}$ gives the bound against linear φ. At $w=0$ the maximizer is the default's act $a_0=(m_0-\chi)/\lambda$, so by the envelope theorem $S'_\varphi(0)=\varphi(\mathbb{E}_ku(a_0,\theta))-\varphi(U_0)$, and (8) gives $\mathbb{E}_ku(a_0,\theta)=U_k-D_k$. Concavity gives $\varphi(U_k-D_k)-\varphi(U_0)\le\Delta U_k-D_k$, and monotonicity gives the sign. So the optimal weight is positive whenever $\Delta U_k>D_k$, and zero whenever the linear DM's is; since linear activation requires $\Delta U_k>T_k(\overline{W}_k)$ with $T_k\le D_k$ (Proposition 1), this gives the band, and the comparison of φ with $\tilde{\varphi}$ gives the last claim.

(ii) Let $\delta := m_k - m_0$, so that $\Delta U_k = O(\delta)$ and $D_k = O(\delta^2)$, and let $\bar{U} := (1-w)U_0 + wU_k$. Expanding φ to second order around $\bar{U}$, the first-order terms cancel and $(1-w)(U_0-\bar{U})^2 + w(U_k-\bar{U})^2 = w(1-w)(\Delta U_k)^2$, so $(1-w)\varphi(U_0) + w\varphi(U_k) = \varphi(\bar{U}) - \frac{\xi}{2}w(1-w)(\Delta U_k)^2 + O(\delta^3)$. At the linear act $\bar{a} = (\bar{m}-\chi)/\lambda$, $\mathbb{E}_j u(\bar{a},\theta) = U_j - (m_j - \bar{m})^2/(2\lambda)$, and since $\varphi'(U_j) = 1 + O(\delta)$, the single-act value at $\bar{a}$ falls short of $(1-w)\varphi(U_0) + w\varphi(U_k)$ by $w(1-w)D_k + O(\delta^3)$; re-optimizing the act gains only $O(\delta^4)$, since the gradient at $\bar{a}$ is $O(\delta^2)$. Hence $S_\varphi(w) = (1-w)\varphi(U_0) + w\varphi(U_k) - w(1-w)D_k + O(\delta^3)$, and since $\varphi(U_k) - \varphi(U_0) = \Delta U_k - \frac{\xi}{2}(\Delta U_k)^2 + O(\delta^3)$, the gain from holding k is $w\Delta\tilde{U}_k - w(1-w)D_k - \frac{\gamma}{2}w^2 + O(\delta^3)$, which is (12) with $\Delta\tilde{U}_k$ in place of ΔU_k. The ambiguity premium lowers $(1-w)\varphi(U_0) + w\varphi(U_k)$ below $\varphi(\bar{U})$ but adds no curvature in w: the sum is affine in w, so the premium's concavity is exactly offset by that of $\varphi(\bar{U})$. □

Proof of Proposition 6. (i) By definition $G = \mathrm{CE} - H$ and $\mathrm{CE} = \bar{U} - \pi$, and likewise after the signals; subtracting and taking expectations gives the decomposition. Under linear φ, $\mathrm{CE} = \bar{U}$, $\pi = 0$, and $G = \bar{U} - A(\mathbf{w})$ by (7), so $H = A$. Since $\mathbb{E}_{\mathbf{w}}[w_{k,n} g(y^n)] = w_k \mathbb{E}_k[g(y^n)]$ for any g, $\mathbb{E}_{\mathbf{w}}[\bar{U}_n] = \sum_k w_k \mathbb{E}_k[(m_{k,n}-\chi)^2]/(2\lambda) = \bar{U} + \sum_k w_k \psi_k(n)/(2\lambda)$, and the decomposition is (14).

(ii) A revealing signal sends $w_{k,n}$ to one with probability w_k and to zero otherwise and leaves the predictions unchanged, so $\mathbb{E}_{\mathbf{w}}[\bar{U}_n] = \bar{U}$ and $\pi_n = H_n = 0$, which gives $H + \pi$. If the models recommend the same act a^*, then $\mathbb{E}_j u(a^*,\theta) = U_j$, so $G = \mathrm{CE}$ and $H = 0$. As φ^{-1} of a nonconstant affine function, CE is strictly convex in w_k, so π is strictly concave, with $\pi = 0$ at $w_k \in \{0,1\}$. Holding U_0 fixed, $\partial\pi/\partial U_k = w_k[1 - \varphi'(U_k)/\varphi'(\mathrm{CE})]$, which has the sign of $U_k - \mathrm{CE}$ and hence of ΔU_k, since CE lies strictly between U_0 and U_k. If $\tilde{\varphi} = h \circ \varphi$ with h increasing and concave, Jensen's inequality gives $\widetilde{\mathrm{CE}} \le \mathrm{CE}$, so $\tilde{\pi} \ge \pi$. The second-order form follows from the expansion in the proof of Proposition 5(ii), which gives $\mathrm{CE} = \bar{U} - \frac{\xi}{2}w_k(1-w_k)(\Delta U_k)^2 + O(\delta^3)$. □

Endogenous exposure (Section 3.6).

Proof of Proposition 7. Under model j, $y = d'\tilde{x} + \varepsilon$ has predictive mean $\mu_j = d' b_j m_j$ and predictive variance $s_j^2 = d'\Omega_j d + \nu_y^2$, with $\Omega_j := \Sigma + b_j b_j'/P_j(0)$; write $s^2 := s_0^2$. Because $s_k^2 - s^2 = d'(\Omega_k - \Omega_0)d$ does not grow with ν_y^2, replacing s_k^2 by s^2 below changes the posterior responses only by terms of relative order $1/\nu_y^2$, which do not affect the leading term. After observing y, model j's posterior mean is $m_j + g_j(y - \mu_j)$ with $g_j = (d' b_j/P_j(0))/s^2$, and the updated weight $w_1 = wf_k/(wf_k + (1-w)f_0)$ satisfies $\partial w_1/\partial y = w_1(1-w_1)\,d'\delta/s^2$, since $\mu_k - \mu_0 = d'\delta$. With quadratic payoffs and linear φ the act tracks the mixture mean $\bar{m}(y) = (1 - w_1)[m_0 + g_0(y-\mu_0)] + w_1[m_k + g_k(y-\mu_k)]$, and the value of the observation is $\mathrm{Var}(\bar{m}(y))/(2\lambda)$. For a weak observation, linearize $\bar{m}$ around the prior:

$$\frac{\partial \bar{m}}{\partial y} \;=\; \frac{(1-w)\,d'b_0/P_0(0) + w\,d'b_k/P_k(0) + w(1-w)(m_k - m_0)\,d'\delta}{s^2} \;=\; \frac{d'v(w)}{s^2},$$

so the value is $(d'v(w))^2/(2\lambda s^2)$ to leading order as $\nu_y^2 \to \infty$, the remainder being of smaller order because the posterior responses g_j and $\partial w_1/\partial y$ are $O(1/s^2)$. Under the normalization $d'\Omega_0 d = 1$, s^2 is constant, and the Cauchy–Schwarz inequality in the Ω_0 inner product maximizes $(d'v)^2$ at $d \propto \Omega_0^{-1} v(w)$. Part (i) follows by setting $w \in \{0,1\}$ in $v(w)$. For part (ii), the coefficient on δ in $v(w)$ is $w(1-w)(m_k - m_0)$, which vanishes when $m_k = m_0$. That d^* most reduces model uncertainty is a separate optimization from the one that defines $d^o(w)$. By the martingale property of the weight (Lemma OA.2), $\mathbb{E}[w_1(1-w_1)] = w(1-w) - \mathrm{Var}(w_1)$ exactly, and for a weak observation $\mathrm{Var}(w_1) = w^2(1-w)^2(d'\delta)^2/s^2$ to leading order, which the normalization maximizes at d^*. The direction $d^o(w)$ instead maximizes the whole value of the observation, which also rewards refining each model through $b_0/P_0(0)$ and $b_k/P_k(0)$. The two coincide only when these refinement terms are proportional to δ; otherwise $d^o(w)$ gives up some reduction in model uncertainty for refinement. □

Context effects (Section 4.1).

Proof of Corollary 2. By Lemma 1, $s_k = \alpha_k' x$ with $\alpha_k \propto \Sigma^{-1} b_k$, so $s_k(x+\zeta) = s_k(x) + b_k'\Sigma^{-1}\zeta/h_k$. In (13), $R_0 - R_k = h_k s_k^2 - h_0 s_0^2$ because $x'\Sigma^{-1}x$ cancels, q_0 and q_k depend on x only through s_0 and s_k, and c_0, c_k do not depend on x. Hence $\log\Lambda_k$ depends on x only through (s_0, s_k), and $s_0(x+\zeta) = s_0(x)$ because ζ is payoff-irrelevant under the default. (i) If $b_k'\Sigma^{-1}\zeta = 0$ for every k, no s_k changes, so no Λ_k and, by (5), no $\overline{W}_k$ changes; since each model's posterior depends on x only through s_k, the values U_k, the disagreements D_k, the objective (4), and its constraints are unchanged as functions of $(\mathbf{w}, n)$, and so is the optimum. (ii) Λ_k changes only if s_k does, that is, only if $b_k'\Sigma^{-1}\zeta \neq 0$; for an alternative j that does not read ζ, $\overline{W}_j = \Lambda_j/(1+\sum_i \Lambda_i)$ changes only through the denominator. □

Proof of Corollary 3. Let f_k and f_0 be the two predictive densities of y along d and let y be generated by f. The expected increment of $\log\Lambda_k$ is $\mathbb{E}_f[\log f_k(y) - \log f_0(y)] = \mathrm{KL}(f\,\|\,f_0) - \mathrm{KL}(f\,\|\,f_k)$. If $f = f_k$ this is $\mathrm{KL}(f_k\,\|\,f_0) \geq 0$, which for Gaussians equals $(\mu_k - \mu_0)^2/(2s_0^2) + \frac{1}{2}[\log(s_0^2/s_k^2) + s_k^2/s_0^2 - 1]$ and is increasing in the squared mean difference at fixed variances; if $f = f_0$ it is $-\mathrm{KL}(f_0\,\|\,f_k) \leq 0$. At $w_k = 0$ no signal is acquired (Corollary 1). □

Interference effects (Section 4.2).

Proof of Proposition 8. Fix n and the total weight off the default, $W = \sum_{k\geq 1} w_k$, and let $P_W := \{\mathbf{w} \geq 0 : \sum_{k\geq 1} w_k = W,\ w_k \leq \overline{W}_k(x)\}$. On P_W the objective (4) is strictly convex: with $w_0 = 1 - W$ fixed, the single-act term $\max_a \sum_k w_k \varphi(\mathbb{E}_k u(a,\theta))$ is a maximum of functions affine in $\mathbf{w}$ and hence convex, and the within-model term is affine in $\mathbf{w}$; and by the identity in the proof of Lemma 2, $-C(\mathbf{w}) = -\frac{\eta}{2}W^2 + \frac{\eta-\gamma}{2}\sum_k w_k^2$, which is strictly convex on the affine set $\{\sum_k w_k = W\}$ because $\eta > \gamma$. A strictly convex function attains its maximum over a polytope only at extreme points. A point of P_W with two coordinates $w_i, w_j \in (0, \overline{W})$ is not extreme, since $\mathbf{w} \pm \varepsilon(e_i - e_j) \in P_W$ for small ε. Hence every maximizer on P_W has at most one coordinate strictly between zero and its cap. An optimum of the full problem maximizes (4) on P_{W^*} for its own total W^*, so the same holds for it. If no cap binds, the only extreme points of P_W are the vertices We_k, so at most one alternative is held. □

Proof of Proposition 9. (i) For $j \neq k$, $\partial\overline{W}_k/\partial L_j = -L_k/(L_0 + \sum_{i\geq 1} L_i)^2 < 0$; setting $L_j = 0$ for all $j \neq k$ gives the comparison with $L_k/(L_0 + L_k)$. (ii) With $J = \{k\}$ the bound is $e^\beta L_k/(L_0 + e^\beta L_k + \sum_{j\neq k} L_j)$; with $k \in J$ and $|J| \geq 2$ the denominator is larger by $(e^\beta - 1)\sum_{j\in J\setminus\{k\}} L_j > 0$. (iii) Write $d_k := m_k - m_0$. By (11), the hedging cost is $\sum_k w_k D_k - (\sum_k w_k d_k)^2/(2\lambda)$, and with (3),

$$V(\mathbf{w}, n) - V(0, n) \;=\; \sum_{k\geq 1} f_k(w_k) + \sum_{j<k} w_j w_k \big[d_j d_k/\lambda - \eta\big],$$

where $f_k(w) := w(\Delta U_k - D_k) - w^2(\gamma/2 - D_k)$ is the gain from holding k alone. In the all-or-nothing regime f_k is convex with $f_k(0) = 0$ and $f_k(\overline{W}_k) = \overline{W}_k[\Delta U_k - T_k(\overline{W}_k)] < 0$, so $f_k < 0$ on $(0, \overline{W}_k]$; with every pairwise term nonpositive, $V(\mathbf{w}, n) < V(0, n)$ for every feasible $\mathbf{w} \neq 0$. Hence $\mathbf{w}^* = 0$, and $n^* = 0$ under Assumption 1 (Proposition 2). In the graded regime, at any feasible weights the value of n signals is $\sum_k w_k \psi_k(n)/(2\lambda) \leq cn\sum_k w_k/\underline{w}_k \leq cn\sum_k \overline{W}_k/\underline{w}_k$, since $\psi_k(n) \leq n\psi_k(1)$ and $\psi_k(1)/(2\lambda) = c/\underline{w}_k$, which does not exceed the cost cn when $\sum_k \overline{W}_k/\underline{w}_k \leq 1$. □

Proof of Corollary 4. Holding j alone at weight w_j gives, by (12), $V(w_j) - V(0) = f_j(w_j) := w_j(\Delta U_j - D_j) - w_j^2(\gamma/2 - D_j)$. If $\gamma \geq 2D_j$, $f_j(w) \leq w(\Delta U_j - D_j)$; if $\gamma < 2D_j$, f_j is convex with $f_j(0) = 0$, so $f_j(w) \leq w f_j(1) = w(\Delta U_j - \gamma/2)$. Either way $f_j(w_j) \leq -\delta_j w_j$ with $\delta_j := \min\{D_j, \gamma/2\} - \Delta U_j > 0$, so j is never activated on its own, whatever its bound.

Without the cap, adding j only enlarges the feasible set, so if an alternative was held before, the optimum still exceeds $V(0)$, which holding j alone cannot, and some alternative other than j is held.

With the cap, write $\Lambda := \sum_{i \neq j} \Lambda_i$. Adding j scales every other bound by $(1+\Lambda)/(1+\Lambda+\Lambda_j) \to 0$ as $\Lambda_j \to \infty$. Take a menu with one other alternative k in the all-or-nothing regime with $T_k(\overline{W}_k) < \Delta U_k < D_k$, so that k is activated before j is added, and let ε be k's bound after j is added. By the decomposition in the proof of Proposition 9(iii), $V(w_k, w_j) - V(0,0) = f_k(w_k) + f_j(w_j) + w_k w_j [d_k d_j/\lambda - \eta]$. Since $|d_k d_j|/\lambda = 2\sqrt{D_k D_j}$ and $w_k \leq \varepsilon$, the last term is at most $2\sqrt{D_k D_j}\,\varepsilon\, w_j$, and in the all-or-nothing regime $f_k(w_k) \leq w_k[\Delta U_k - D_k + (D_k - \gamma/2)\varepsilon]$. Hence $V(w_k, w_j) - V(0,0) \leq w_k[\Delta U_k - D_k + (D_k - \frac{\gamma}{2})\varepsilon] + w_j[2\sqrt{D_k D_j}\,\varepsilon - \delta_j]$, negative unless $w_k = w_j = 0$ once ε is small, as a large Λ_j ensures. Then $\mathbf{w}^* = 0$ and, under Assumption 1, $n^* = 0$ (Proposition 2). Imposing the inequalities at their suprema over n makes ε uniform in n. □

Templates and mental models (Section 5.1).

Proof of Lemma 3. With $S = \{t\}$ and one situation, $\omega_t + e_{t,\ell} \sim \mathcal{N}(0, \sigma_t^2)$ whatever ρ_t, so $\vartheta_{t,\ell} \sim \mathcal{N}(\mu_t, \sigma_t^2)$ and (15) is (1). For a general node, x is Gaussian with mean $B_S \mu_S$ and covariance $V_S = \Sigma + B_S \Sigma_S B_S'$ in each situation, so its marginal likelihood is a Gaussian density exactly as in Lemma OA.3; (5), (3), (7), and (4) are defined from marginal likelihoods, weights, and model-contingent values, none of which refers to the internal structure of the model. □

Vertical creative thinking (Section 5.2).

Proof of Proposition 10. Under S, $x \sim \mathcal{N}(B_S \mu_S, V_S)$; under $S \cup \{t\}$, $x \sim \mathcal{N}(B_S \mu_S + b_t \mu_t,\ V_S + \sigma_t^2 b_t b_t')$, the template's parameter being independent of the others. With $\mu_t = 0$, the Sherman–Morrison formula and the matrix determinant lemma give the log Bayes factor

$$\log \Lambda_{S \cup t : S} \;=\; \tfrac{1}{2}\Big[\frac{\sigma_t^2 (b_t' V_S^{-1} r)^2}{1 + \sigma_t^2 h_t} - \log(1 + \sigma_t^2 h_t)\Big] \;=\; \tfrac{1}{2}\big[G_t - \log(1 + \sigma_t^2 h_t)\big],$$

with $G_t = \sigma_t^2 h_t z_t^2/(1 + \sigma_t^2 h_t)$, which is (17). In the two-node comparison the step is admitted if and only if the log Bayes factor is positive, which is (18). Writing $u := \sigma_t^2 h_t$ and $f(u) := (1+u)\log(1+u)/u$, we have $f'(u) = [u - \log(1+u)]/u^2 > 0$ and $f(u) \to 1$ as $u \to 0$. □

Horizontal creative thinking (Section 5.3).

Proof of Proposition 11. Under the local node the d_ℓ are independent $\mathcal{N}(0, 1 + \sigma^2)$; under the general node at $\rho = 1$, $d \sim \mathcal{N}(0, I + \sigma^2 \mathbf{1}\mathbf{1}')$, with determinant $1 + N\sigma^2$ and quadratic form $SS + N\bar{d}^2/(1 + N\sigma^2)$, since $\sum_\ell d_\ell^2 = SS + N\bar{d}^2$. Subtracting the local log likelihood gives (19), which vanishes at $N = 1$ ($SS = 0$). For $N \geq 2$ it is affine, decreasing in SS and increasing in $\bar{d}^2$, which gives the threshold, and the Occam gain rises with N by $\frac{1}{2}[\log(1+\sigma^2) - \log(1 + \sigma^2/(1+N\sigma^2))] > 0$. □

Deeper explanations apply more widely (Section 5.4).

Proof of Proposition 12. (i) On R_t, $\rho_t = 1$ sets $e_{t,\ell} = 0$ in (15), so $\vartheta_{t,\ell} = \mu_t + \omega_t$ is common to every situation in R_t; outside R_t the template does not contribute and the node coincides with S. The reach is therefore R_t. (ii) By (i), transfer over R_t is evaluated at $\rho = 1$, and at $SS = \bar{d} = 0$ the Bayes factor reduces to the Occam term, which is strictly increasing in ρ for $N \geq 2$ (proof of Proposition OA.7). (iii) The dispersion penalty $\frac{1}{2} SS[1/a - 1/(v + \sigma^2)]$ is increasing in ρ because a is decreasing in ρ, and is largest at $\rho = 1$. □

Critical thinking (Section 6).

Proof of Proposition 13. Write $m := |\mathscr{S}|$. When x is generated by S, $r \sim \mathcal{N}(0, V_S)$, so each $z_t = b_t' V_S^{-1} r/\sqrt{h_t}$ is standard normal, and $\mathrm{Cov}(z_t, z_{t'}) = b_t' V_S^{-1} b_{t'}/\sqrt{h_t h_{t'}} = 0$ for $t \neq t'$ by orthogonality in the V_S^{-1} metric; the

z_t are therefore independent. An environment with d observable features admits at most d such directions, so the limits in m are along a sequence of environments with $d \geq m$; part (ii) uses neither orthogonality nor this growth. By Proposition 10 with common υ, $\log \Lambda_{S\cup t:S} = \frac{1}{2}[\frac{\upsilon}{1+\upsilon} z_t^2 - \log(1+\upsilon)]$ is increasing in z_t^2, so t^* maximizes z_t^2, and the discounted log odds (20) are positive if and only if $\max_t z_t^2 > c_\kappa(\upsilon, m)$, which is (21). Set $\beta := \kappa(1+\upsilon)/\upsilon$, so that $c_\kappa = \frac{1+\upsilon}{\upsilon}\log(1+\upsilon) + 2\beta \log m$ and $e^{-c_\kappa/2} \propto m^{-\beta}$.

(i) If $\kappa < \upsilon/(1+\upsilon)$ then $\beta < 1$. With $p := \Pr(|Z| > \sqrt{c_\kappa})$ for $Z \sim \mathcal{N}(0,1)$, independence gives $1 - \Pi_\kappa(m) = (1-p)^m \leq e^{-mp}$, and the Mills-ratio bound $\Pr(|Z| > u) \geq \sqrt{2/\pi}\,\frac{u}{1+u^2} e^{-u^2/2}$ gives $mp \geq \text{const} \cdot m^{1-\beta}\sqrt{c_\kappa}/(1+c_\kappa) \to \infty$, since c_κ grows only logarithmically in m. Hence $\Pi_\kappa(m) \to 1$; the naive DM is the case $\kappa = 0$.

(ii) The union bound, which needs only that each z_t is standard normal, and the tail bound $\Pr(|Z| > u) \leq e^{-u^2/2}$ give $\Pi_\kappa(m) \leq m e^{-c_\kappa/2} \propto m^{1-\beta} \to 0$ when $\beta > 1$, that is, when $\kappa > \upsilon/(1+\upsilon)$.

(iii) A candidate is admitted if and only if $z_t^2 > c_\kappa(\upsilon, m)$, and $c_\kappa - c_0 = 2\kappa\frac{1+\upsilon}{\upsilon}\log m$. □

Proof of Corollary 5. For each situation ℓ, split the residual r_ℓ into its component along t, $u_\ell := b_t' V_S^{-1} r_\ell / h_t$, and a remainder that is uncorrelated with u_ℓ, jointly Gaussian with it, and identically distributed under both nodes. Under S the u_ℓ are independent $\mathcal{N}(0, 1/h_t)$; under $S\cup\{t\}$, $u_\ell = \vartheta_t + e_\ell$ with $\vartheta_t \sim \mathcal{N}(0, \sigma_t^2)$ common to all situations and e_ℓ independent $\mathcal{N}(0, 1/h_t)$. The contrasts $u_\ell - \bar{u}$ have the same distribution under both nodes and are independent of $\bar{u} := N^{-1}\sum_\ell u_\ell$, so the Bayes factor is the ratio of the densities of $\bar{u}$, $\mathcal{N}(0, \sigma_t^2 + 1/(Nh_t))$ against $\mathcal{N}(0, 1/(Nh_t))$. With $Z_t := \sqrt{Nh_t}\,\bar{u} = N^{-1/2}\sum_\ell z_{t,\ell}$ and $\upsilon = \sigma_t^2 h_t$,

$$\log \Lambda_{S\cup t:S} \;=\; \tfrac{1}{2}\Big[\frac{N\upsilon}{1+N\upsilon} Z_t^2 - \log(1+N\upsilon)\Big],$$

the expression of Proposition 10 with υ replaced by $N\upsilon$ and z_t by Z_t.

(i) As in the proof of Proposition 13, the discounted log odds are positive if and only if $Z_t^2 > c_\kappa(N\upsilon, |\mathscr{S}|)$. (ii) The difference follows from (21): it is zero at $|\mathscr{S}| = 1$, increasing in $|\mathscr{S}|$, and decreasing in N because $(1+u)/u$ is decreasing in u. (iii) If t is correct with parameter ϑ_t, then $\bar{u} \sim \mathcal{N}(\vartheta_t, 1/(Nh_t))$, so $Z_t = \sqrt{Nh_t}\,\vartheta_t + \varepsilon_0$ with $\varepsilon_0 \sim \mathcal{N}(0,1)$ and $Z_t^2/N \to h_t\vartheta_t^2 > 0$ in probability, whereas $c_\kappa(N\upsilon, |\mathscr{S}|) \leq (1+1/\upsilon)[\log(1+N\upsilon) + 2\kappa\log|\mathscr{S}|]$ grows only logarithmically in N. Hence $\Pr(Z_t^2 > c_\kappa(N\upsilon, |\mathscr{S}|)) \to 1$. □