BAYESIAN INTELLIGENCE FROM THE OUTSIDE

Anonymous authors
Paper under double-blind review

ABSTRACT

Inferring mechanisms underlying intelligence from observable behavior is a foundational challenge in artificial intelligence. We develop a theory of Bayesian intelligence for agents such as language models, which tests whether observed behavior could be generated by a rational agent responding to information revealed by inference-time reasoning. Each prompt induces a possibly imperfect internal experiment; the agent updates a full-support prior by Bayes' rule and faithfully reports its posterior over the possible answers to the question. Repetitions draw fresh, independent outcomes from the same unobserved experiment at one fixed state. We show that the agent's behavior admits this explanation if and only if its reports are not fully contradictory, i.e., some state remains possible under every report across all prompts. Report frequencies and the sizes of positive probabilities impose no further restrictions. We further propose and characterize the behavioral implications of one agent having access to a more informative experiment: there should exist a coupling of report distributions such that the more informative agent's report excludes every answer excluded by its counterpart. Finally, we show the difficulty of aggregating coarse reports: unless the agent reports a belief about the complete state of the world, the optimal aggregation can assign arbitrary weights to states that have not been excluded. These results provide a basis for understanding when agents' behavior can be modeled as revealing information through reasoning and highlight the difficulty of rejecting Bayesian rationality.

1 INTRODUCTION

As systems based on large language models (LLMs) are integrated deeper into decision-making and economic contexts (Deng et al., 2024; Dütting et al., 2024; 2026; Fish et al., 2026), understanding the mechanism through which LLMs reason and whether this process is internally consistent or rationalizable becomes increasingly important. Internal consistency is a necessary component of valid reasoning and determines whether a system can be modeled from the outside as a rational agent. For example, classic representation theorems provide axiomatic foundations for beliefs as drivers of rational behavior: when behavior is sufficiently consistent, it can be represented by a utility function and subjective probabilities over the relevant uncertainty, such that an agent behaves "as if" maximizing expected utility given those beliefs (Savage, 1954; Anscombe and Aumann, 1963). Bayes' rule (Bayes, 1763) further prescribes how beliefs should be updated in response to external information. A growing literature tests whether LLMs update in a Bayes-consistent manner to evidence supplied in the prompt (Chen et al., 2026; Zhu and Griffiths, 2024; Andrews and Sarkar, 2026), or whether they express preferences that are sufficiently internally coherent to be rationalizable (Mazeika et al., 2026).

In addition to externally provided evidence, sophisticated behavior from LLMs is now underpinned by substantial inference-time reasoning. Our goal is to develop an analogous framework for testing the internal consistency of this process: we model inference-time computation as supplying "internal" information that updates the agent's beliefs. Testing whether the update process is rationalizable in Bayesian terms allows us to separate quality of the information acquired (which may be variable, depending e.g., on the model's domain-specific or factual knowledge) from the correctness of updating.

The challenge is that the information driving LLMs' answers is often opaque: a large body of work questions whether the text of LLM chain of thought provides faithful explanations of the output (Lanham et al., 2023; Chen et al., 2025), or whether reasoning always plays a causal role how LLMs

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

solve problems (Boppana et al., 2026; Palod et al., 2025). We provide a complementary perspective on such problems by asking how much can be tested within a black-box framework. Formally, we introduce a model based on the classical Blackwell framework of statistical experiments where reasoning reveals *information* to the LLM, which the LLM is then able to condition on in producing its final response. We do not make any assumptions about whether this intermediate information is "faithful" when viewed as natural language. Instead, we ask just whether the distribution of final reports from the LLM under different prompts or repeated resampling of the reasoning process is internally rationalizable, i.e., Bayes-consistent. Building on the same framework, we also show how to test whether on agent's reasoning reveals strictly more information than another, and how repeated reports from a LLM should be combined.

Our framework separates the correctness of reasoning, as judged by consistency with external ground truth, from its internal consistency by allowing for the possibility experiment can be imperfect, so the agent can place little posterior probability on the truth and give an incorrect answer. The experiment may also depend arbitrarily on the prompt: our formal model allows the possibility that a change in framing can change the reasoning path. The model does require correct specification in two senses: the prior gives the true state positive probability, and the actual signal distribution is the one used in Bayes' rule. Thus, conditional on its signal, the agent updates without error.

Our first result, Theorem 1, characterizes the collections of *Bayes-plausible* report distributions, i.e., those consistent with this model. A collection is Bayes-plausible if and only if some single complete state remains possible under every report for every prompt. In other words, zero-probability claims—"hard exclusions"—must be jointly consistent; conditional report frequencies and the numerical values of positive probabilities are otherwise unrestricted. In particular, any collection of distributions containing only reports that assign positive probability to every answer is rationalizable. No martingale restriction follows from sampling reports at a fixed state. This rationalization establishes compatibility with Bayesian reasoning, without identifying how the agent actually reasons.

We next compare the information acquired by two agents. For a given prompt, one report distribution *plausibly Blackwell dominates* another if both can arise at a common prior and true state, with a more informative experiment behind the first. When each possible answer is compatible with at least two complete states, Theorem 2 shows that this holds exactly when all reports share a possible true answer and the reports can be paired so that each more informative report rules out every answer ruled out by its less informative counterpart, and possibly more. Thus, the observable restrictions on a plausible intelligence order concern coherent hard exclusions. With only interior reports, either order is compatible with the observations, regardless of confidence or variability.

Finally, we ask what repeated reports reveal about their Bayesian aggregate when the reasoning experiment is unknown. By Theorem 3, with at least two distinct reports and at least two complete states per answer, the aggregate must rule out every answer excluded by a report, but can assign any weights to the other answers. If one report rules out an answer and another does not, their arithmetic average violates this restriction. Affine log-odds pooling respects the restriction, and the optimal rule in the Gaussian intelligence model takes this form. When reports give beliefs about the complete state, a prior-corrected sum of their log odds determines the Bayesian aggregate. This requires knowing the prior, but not the experiment. Coarse reports leave out distinctions between states that are needed for this calculation.

1.1 Related Literature

Bayesian Reasoning in Language Models. One strand of work models in-context learning as implicit Bayesian inference under a specified generative model of pretraining data and prompts (Xie et al., 2022). We ask which report distributions can be explained by Bayesian updating when the experiment induced by each prompt is unknown and unrestricted. The aim is to establish what can be tested from these reports; the model does not explain how pretraining produces Bayesian inference.

Many existing evaluations test several forms of logical or probabilistic consistency. BeliefBank treats categorical answers about propositions as a single, stable set of beliefs (Kassner et al., 2021). Other studies check probability identities across prompts (Zhu and Griffiths, 2024; Imran et al., 2025) or compare beliefs elicited later with Bayesian updates calculated from elicited priors and likelihoods (Chen et al., 2026). Martingale tests examine in-context predictions or reasoning his-

tories (Falck et al., 2024; He et al., 2025), while evidence-based tests vary the informativeness and reliability of supplied evidence (Kim et al., 2025). These evaluations maintain implicit and explicit assumptions about what responses represent and how information is related across prompts. We derive the implications of Bayesian updating when those information relationships are unrestricted. A categorical answer need not express certainty, and different prompts can induce different reasoning experiments.

Other benchmarks compare probabilistic judgments with reference distributions. CogBench's urn task supplies priors and likelihoods and estimates how models weigh them (Coda-Forno et al., 2024); QUITE tests numerical probabilities and verbal uncertainty given verbalized Bayesian networks and evidence (Schrader et al., 2024); and Paruchuri et al. (2024) evaluate percentile estimation, sampling, and probability calculations. These evaluations measure how well models solve particular inference tasks. We ask whether their reports can be explained by Bayesian updating under an unknown internal experiment that may provide imperfect information (see discussion in Section 6 on a possible combination of these approaches).

Bayesian Learning and Its Empirical Content One strand of work studies which distributions of posterior beliefs can arise before the state is known. The splitting lemma of Aumann and Maschler (1995) and the persuasion formulation of Kamenica and Gentzkow (2011) show that a finite distribution of posteriors can be induced by a signal if and only if its mean equals the prior. Arieli and Babichenko (2024) extend this characterization to a homogeneous population. These results average over both states and signals under the prior-predictive distribution. By contrast, Arieli et al. (2020) characterize forecast distributions conditional on a realized state and ask when they identify that state despite an unknown information structure. Similarly, Doval and Smolin (2024) study welfare conditional on each realized type, retaining information hidden by ex ante averaging. We condition on one state, but study the joint feasibility of coarse reports from multiple prompt-dependent experiments, as well as compatible information orders and aggregates. Because our observations do not average across states, Theorem 1 imposes much weaker restrictions.

A second strand asks what Bayesian updating implies when the analyst does not fully know the agent's information structure. Shmaya and Yariv (2016) obtain an "anything goes" result: any assignment of posterior beliefs to signal histories can be explained by Bayesian updating when subjects' conjectures about the joint distribution of the state, signals, and number of signals revealed are unrestricted. Subjects can treat the experimenter's choice of how much information to reveal as evidence about the state, beyond the signals' content. Protocols that restrict these conjectures recover testable implications. Molavi (2024) shows that martingale restrictions can also fail when the agent's subjective signal distribution differs from the objective distribution. Molavi (2026) further shows that unrestricted misspecification on a finite state space leaves only support restrictions. In our setting, the agent's signal model is correctly specified, but the prompt-dependent internal experiments are unobserved. Repeated reports are sampled at one fixed state and must all leave it possible; unobserved signal probabilities at other states supply the remaining freedom for rationalization.

Behavioral and Experimental Economics Our theory can also be applied to human intelligence and therefore speaks to behavioral and experimental economics. Like our model, the tracing procedure of Harsanyi and Selten (1988) formalizes an internal reasoning process that starts from prior beliefs. It instead adjusts beliefs about others' strategies through iterated best-response reasoning until plans and expectations coincide at an equilibrium; our agent conditions on a prompt-induced signal about the state and may remain uncertain. Rational inattention endogenizes the information structure: agents choose what and how much to learn under a finite Shannon capacity (Sims, 2003) or a Shannon information cost that yields generalized multinomial-logit choice (Matějka and McKay, 2015). We instead allow an arbitrary prompt-dependent experiment and ask what Bayesian updating alone implies, without assuming optimal information acquisition or a particular cost. Recently, Caplin et al. (2025) apply cognitive-economic revealed-preference tests to a trained image classifier, varying its training loss and rationalizing its predictions through costly learning. Their object is learning across training environments; ours is posterior formation during inference by a fixed trained system. We note that AI systems are especially useful for this inquiry because an identical system and configuration can be queried repeatedly in a controlled environment. Human subjects differ and remember prior exposure, which confounds repeated measurement and requires more restrictive designs.

2 MODEL

An agent responds to prompts from an observer. We propose a Bayesian theory of how it acquires information, updates its beliefs, and responds.

Formally, we let $\omega \in \Omega$ denote a state of the world, where Ω is finite and $|\Omega| = N$. The state encompasses *all* uncertainty about the world relevant to the agent's answers, so we expect N to be very large. The agent starts with a full-support *prior*

$$\mu_0 \in \operatorname{int} \Delta(\Omega).$$

This prior is the starting point for the agent's computation and can be interpreted as summarizing its knowledge about the world in the absence of any additional computation.

The observer sends a prompt p from a given vocabulary P. Each prompt triggers a reasoning process that leads the agent to revise its belief via computation that includes the analysis of the prompt, the drawing on its internal knowledge, the chain-of-thought procedure (Wei et al., 2022), and possibly a retrieval of additional external information via tools. The agent then outputs a report r from a given vocabulary R.

We model this process as a Bayesian learning about the state of world followed by a strategic response. First, the agent gets access to a prompt-dependent statistical experiment

$$E_p = (S_p, \pi_p), \qquad \pi_p : \Omega \to \Delta(S_p), \tag{1}$$

with finite S_p. The agent observes a signal $s \in S_p$ and, building on its prior, forms the *posterior* by Bayes' rule,

$$\mu_p(s)(\omega) = \frac{\mu_0(\omega)\pi_p(s \mid \omega)}{\sum_{\omega' \in \Omega} \mu_0(\omega')\pi_p(s \mid \omega')},$$

whenever s has positive ex ante probability. Second, the agent responds according to a *strategy* that maps the prompt and posterior to a distribution over responses:

$$\sigma : P \times \Delta(\Omega) \to \Delta(R).$$

We focus on a family of prompts that request the agent's posterior $\nu_p \in \Delta(\mathcal{Y}_p)$ over the possible answers to a question. We represent the correct answer by a state-measurable surjective random variable $Y_p : \Omega \to \mathcal{Y}_p$ and define

$$\nu_p(\mu)(y) = \sum_{\omega \in \Omega : Y_p(\omega) = y} \mu(\omega), \quad \forall\, y \in \mathcal{Y}_p. \tag{2}$$

This focus is without loss of generality as we discuss in Appendix A.1.

For $y \in \mathcal{Y}$, let $\Omega_y := \{\omega' \in \Omega : Y(\omega') = y\}$ denote the set of states compatible with y. In practice, because the state space is enormous and cannot be fully communicated, the partition of Ω induced by Y should be expected to be very coarse. Some of our results impose the weak coarseness assumption $|\Omega_y| \geq 2$ for every $y \in \mathcal{Y}$; we state this assumption explicitly whenever it is needed.

We call the agent *obedient* at prompt p if it reports the requested function with probability one:

$$\sigma(p, \mu) = \delta(\nu_p(\mu)), \quad \forall \mu \in \Delta(\Omega).$$

Here $\delta(x)$ denotes the point mass at x.

Assumption 1 (Obedience)**.** The agent is obedient at every prompt.

The obedience assumption allows us to isolate the restrictions imposed by Bayesian reasoning from errors or strategic distortions in the reporting stage. It strengthens our findings that Bayesian reasoning is difficult to reject.

The observer can submit any prompt to the agent any number of times. The state remains fixed across repetitions, and the experiment associated with the same prompt remains unchanged. As such, the agent's inner workings are stationary and outside the observer's control. However, the

reasoning process begins afresh in each repetition. That is, conditional on the state, the signals are independent and identically distributed:

$$s_1(p), s_2(p), \cdots \mid \omega \overset{\text{i.i.d.}}{\sim} \pi_p(\cdot \mid \omega).$$

If the conditional signal distribution $\pi_p(\cdot \mid \omega)$ is degenerate, i.e., concentrated on a single signal, an obedient agent gives the same answer in every repetition. Otherwise, the agent may form different posteriors and therefore give different reports across repetitions. Thus, Bayesian reasoning is fully compatible with stochastic responses from the agent.

Denote the report distribution induced by prompt p and the true state by

$$\tau_p \in \Delta(\Delta(\mathcal{Y}_p)).$$

Since the signal space is finite, $M_p \triangleq \operatorname{supp} \tau_p$ is finite. By the strong law of large numbers, the empirical distribution of reports converges almost surely to τ_p. An idealized observer who can repeat the prompt indefinitely therefore observes this distribution exactly. Accordingly, we pose our formal results in the following sections in terms of the report distributions.

Remark 1 (Correct Specification)**.** We note that the model implies two forms of correct specification. First, it has a grain of truth: because the prior has full support, the true state receives positive prior probability. Second, the signal distribution is correctly specified: for each prompt p, conditional on each state ω, the signal is drawn from the same distribution $\pi_p(\cdot \mid \omega)$ that the agent uses in Bayes' rule. ♢

Remark 2 (Prompt Dependence)**.** The model does not impose any restrictions on how the agent's reasoning, E_p, depends on the prompt p. Even a small change in the prompt's framing can change responses (e.g., Sclar et al. (2024)). Thus, we isolate the implications of Bayesian reasoning from those of frame invariance. Section 6 discusses additional restrictions linking prompts and their implications for evaluation. ♢

3 Plausible Report Distributions

We first ask when repeated reports can rule out the Bayesian intelligence model. Consider responses to a finite collection of prompts P. Although the prompt-specific experiments may differ, they must share the same prior and true state. We say that a collection of observed report distributions $\{\tau_p\}_{p\in P}$ is *Bayes-plausible* if there exist a prior $\mu_0 \in \operatorname{int} \Delta(\Omega)$, a state $\omega \in \Omega$, and finite experiments $\{E_p\}_{p\in P} = \{(S_p, \pi_p)\}_{p\in P}$ such that, when the true state is ω, the distribution of the agent's posterior report after prompt p is τ_p for every $p \in P$.

Theorem 1 (Bayes-Plausibility)**.** *The collection of report distributions $\{\tau_p\}_{p\in P}$ is Bayes-plausible if and only if*

$$\bigcap_{p\in P} Y_p^{-1}\left(\bigcap_{\nu\in\operatorname{supp}\tau_p} \operatorname{supp}\nu\right) \neq \varnothing. \tag{3}$$

Proof. See Appendix A.3. □

At the true state, a report can only come from a signal with positive probability there. With a full-support prior, Bayes' rule then gives the true answer positive posterior probability. This proves necessity. For sufficiency, the case of a single state is immediate. Otherwise, choose any state consistent with all reports and assign it a sufficiently small positive probability under a common full-support prior. For each prompt, we can extend its answer-level reports to full-state posteriors while preserving their reported marginals. The prior can then be chosen small enough at the candidate state to be able to construct a prompt-specific experiment inducing each distribution at the same state and under the same prior.

The restriction in Theorem 1 is therefore a consistency requirement on hard exclusions across prompts and reports. Positive probabilities and their frequencies impose no further restrictions. For repetitions of one prompt, the condition reduces to $\bigcap_{\nu\in\operatorname{supp}\tau} \operatorname{supp}\nu \neq \varnothing$.

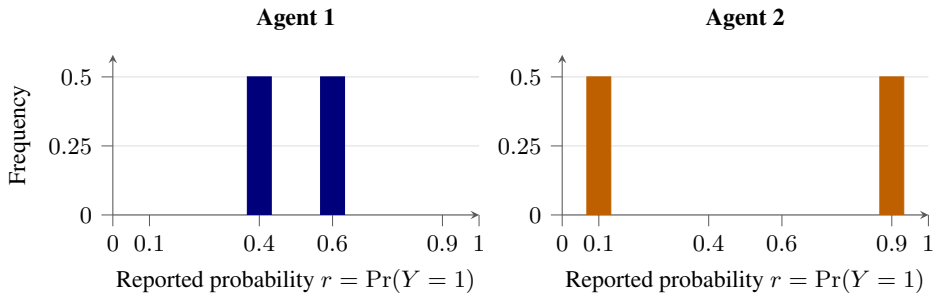


Figure 1: Report distributions for the running example. Each displayed report occurs with probability $1/2$. Agent 1's reports are less variable and less confident than agent 2's.

Remark 3 (Adaptive prompting)**.** This conclusion extends directly to adaptive prompting. Suppose the observer dynamically chooses a sequence of prompts over a finite number of rounds. We can treat each history together with its next prompt as a distinct prompt. By Theorem 1, the whole response tree is Bayes-plausible if and only if a single state ω is compatible with every report at every reachable node. In particular, any such tree of full-support reports is Bayes-plausible. For example, no martingale properties should be expected. ◊

Example 1. Let $r = \Pr(Y = 1)$ denote a report about a binary question, e.g., whether a patient with a given medical history has chronic kidney disease. Figure 1 shows two stylized distributions of repeated reports, $\tau_1 = \frac{1}{2}\delta_{0.4} + \frac{1}{2}\delta_{0.6}$ and $\tau_2 = \frac{1}{2}\delta_{0.1} + \frac{1}{2}\delta_{0.9}$. Both have mean $1/2$, but agent 2's reports are more confident and less consistent. By Theorem 1, both distributions are Bayes-plausible: even the sharp disagreement between agent 2's reports does not rule out Bayesian reasoning. One supporting construction is $\Omega = \{0, 1\}$ with $Y(\omega) = \omega$, common prior $\mu_0 = (5/6, 1/6)$, and true state 1. For rows ordered as $(0, 1)$, the conditional signal probabilities

$$\pi_1 = \begin{pmatrix} 3/20 & 1/15 & 47/60 \\ 1/2 & 1/2 & 0 \end{pmatrix}, \qquad \pi_2 = \begin{pmatrix} 9/10 & 1/90 & 4/45 \\ 1/2 & 1/2 & 0 \end{pmatrix} \tag{4}$$

induce reports $(0.4, 0.6, 0)$ and $(0.1, 0.9, 0)$, respectively. The third signal, although it never occurs at the true state, is required for construction at this prior.

Allowing the prior and experiments to vary, replacing agent 2's lower report by any $\varepsilon > 0$ sufficiently close to zero preserves Bayes-plausibility. Setting it to exactly zero instead forces the candidate true answer to be 0; moving the upper report to 1 as well makes the reports jointly incompatible. The logical implication changes at zero, however small the preceding positive probability was. ⧫

4 Plausible Intelligence Order

Compatibility with Bayesian reasoning does not by itself rank agents. We next ask when their report distributions are compatible with one agent acquiring more informative reasoning outcomes than another. Blackwell informativeness provides a task-specific notion of an intelligence order within our model: it compares the information available for decisions after reasoning, holding prior knowledge fixed.

Formally, we consider two agents and fix a prompt p representing a specific task, so we omit prompt-dependence from notation. The agents have the same prior, but agent $i \in \{1, 2\}$ has access to the experiment $E_i = (S_i, \pi_i)$. Denote by τ_i the report distribution of agent i and let $M_i := \operatorname{supp} \tau_i$.

As standard, experiment E_2 is *Blackwell more informative* than E_1 if E_1 can be generated by first observing the signal from E_2 and then garbling it, i.e., if there exists a stochastic matrix, $G : S_2 \to \Delta(S_1)$, such that for every state $\omega' \in \Omega$ and every signal $s_1 \in S_1$,

$$\pi_1(s_1 \mid \omega') = \sum_{s_2 \in S_2} \pi_2(s_2 \mid \omega') G(s_1 \mid s_2).$$

By Blackwell's theorem (Blackwell (1953)), having access to a more informative experiment weakly improves optimal expected payoffs, evaluated under the prior, in every decision problem.

We say that τ_2 *plausibly Blackwell dominates* τ_1 if there exist a prior $\mu_0 \in \operatorname{int}\Delta(\Omega)$, a true state $\omega \in \Omega$, and experiments E_1, E_2 such that E_2 Blackwell dominates E_1 and the report distribution under E_i, conditional on ω, is τ_i for $i \in \{1, 2\}$.

With a common prior, the posterior distribution generated by the more informative experiment, averaged across states, is a mean-preserving spread of the distribution generated by the less informative experiment. Repeated prompting instead samples reports at one fixed state. The characterization below shows how much this change in the sampling scheme weakens the observable restrictions.

Theorem 2 (Plausible Intelligence Order)**.** *Suppose* $|\Omega_y| \geq 2$ *for every* $y \in \mathcal{Y}$*. The distribution* τ_2 *plausibly Blackwell dominates* τ_1 *if and only if* (i) *there exists* $y \in \bigcap_{\nu \in M_1 \cup M_2} \operatorname{supp}\nu$*, and* (ii) *there exists a coupling* γ *of* τ_2 *and* τ_1 *satisfying* $\gamma(\nu_2, \nu_1) > 0 \Longrightarrow \operatorname{supp}\nu_2 \subseteq \operatorname{supp}\nu_1$.[1]

Proof. See Appendix A.4. □

Intuitively, every report observed at the true state must assign positive probability to the same true value y, which gives (i). Along any report pair with positive coupling mass, a less informative report may restore values that its more informative precursor rules out, but it cannot rule out a value that the precursor assigns positive probability; this is the support inclusion in (ii). Conversely, because the true answer cell contains another state, that state provides enough probability slack to implement any report values consistent with these two support requirements. (If an answer cell is a singleton, there is an additional condition presented in Lemma 2 in the Appendix.) Thus, the plausibility of the intelligence order can be checked as a finite linear feasibility problem on $M_2 \times M_1$, rather than by searching over state-contingent experiments and garblings.

The coupling requirement in Theorem 2 admits an equivalent geometric representation. Associate each report ν with its minimal simplex face $\Delta(\operatorname{supp}\nu)$. Every inclusion-upward-closed collection of faces must then receive weakly more mass under τ_1 than under τ_2. Equivalently, τ_1's induced face distribution first-order stochastically dominates τ_2's under face inclusion. This representation makes it especially clear that only supports of reported beliefs matter for plausible Blackwell dominance.

Conditions for a plausible intelligence order become especially simple when the reported belief is about a binary event, i.e., $|\mathcal{Y}| = 2$. In this case the extreme beliefs are $\nu = (1, 0), (0, 1)$.

Corollary 1 (Intelligence Order. Binary Event.)**.** *If* $|\mathcal{Y}| = 2$ *and* $|\Omega_y| \geq 2$ *for every* $y \in \mathcal{Y}$*, then the distribution* τ_2 *plausibly Blackwell dominates* τ_1 *if and only if (i)* $\operatorname{supp}\tau_1 \cup \operatorname{supp}\tau_2$ *contains at most one extreme belief and (ii)* $\tau_2(\nu) \geq \tau_1(\nu)$ *for both extreme beliefs* ν*.*

Proof. See Appendix A.6. □

Remark 4 (Different Priors)**.** In defining the plausible intelligence order, we required the agents to share a prior to place them on equal footing. However, because the order effectively depends only on the supports of the reports, exactly the same conditions apply even when the agents are allowed to have different priors, e.g., if they are trained on different datasets (see Appendix A.5). ◊

Remark 5 (Evaluation Caution)**.** Plausible dominance means that some compatible experiments have the proposed order; it does not establish that the agents' actual experiments do. Its interpretation also relies on obedience (Assumption 1). For example, under the conditions of Corollary 1, a disobedient agent that mechanically reports the same extreme belief has a distribution that plausibly Blackwell dominates any distribution containing only interior reports. This is a compatible explanation of its outputs, even though the reporting rule itself acquires no information. The criterion therefore restricts possible intelligence orders without providing a sufficient test of actual superiority. ◊

Example 1 (ctd.)**.** On the two-state space of Section 3, no Blackwell ordering can generate the two histograms in Figure 1, under any common prior (see Appendix B.1). Allowing unreported distinctions within an answer makes either order possible. For example, let $\Omega = \{0, 1, 2\}$, with $Y(0) = 0$ and $Y(1) = Y(2) = 1$, true state 1, and the same full-state prior $\mu_0 = (100, 1, 19)/120$ for both agents. This preserves $\Pr(Y = 1) = 1/6$. States 1 and 2 represent circumstances that can

[1] A *coupling* of τ_2 and τ_1 is a distribution $\gamma \in \Delta(M_2 \times M_1)$ with respective marginals τ_2 and τ_1.

affect reasoning without changing the answer to the question. Let E_2 have four signals $s_1, \ldots, s_4$ and conditional signal probabilities (rows ordered as $(0, 1, 2)$)

$$\pi_2 = \begin{pmatrix} 9/200 & 1/200 & 3/20 & 4/5 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 4/19 & 15/19 & 0 \end{pmatrix}. \tag{5}$$

The four signals induce reports $(0.1, 0.9, 0.5, 0)$, respectively. Construct E_1 by garbling s_1 to t_1, s_2 to t_2, and s_4 to t_3; after s_3, choose t_1 or t_2 with equal probability. The resulting signals (t_1, t_2, t_3) induce reports $(0.4, 0.6, 0)$. At true state 1, only the first two signals of each experiment occur, each with probability $1/2$, reproducing the plotted histograms. Thus agent 2's less consistent report distribution is compatible with a strictly more informative experiment. Appendix B.1 constructs the opposite strict order under the same full-state prior. The histograms therefore cannot distinguish these opposite information rankings.

By the proof of Corollary 1, either ranking remains plausible as agent 2's upper report approaches 1, allowing the common prior and experiments to vary. At 1, however, only agent 2 can plausibly be more informative: it sometimes rules out $Y = 0$, while agent 1 excludes neither answer. ♦

5 Plausible Information Aggregation

Repeated reports can also be combined to improve an answer, as in self-consistency methods that aggregate sampled reasoning paths by their final answers (Wang et al., 2023). We ask which aggregate posteriors are compatible with Bayesian updating when the experiment generating the reports is unknown. This connects optimal information extraction from one reasoning agent to robust forecast aggregation across Bayesian experts with an unknown information structure (Arieli et al., 2018).

Formally, the observer has a finite history of T reports $h_T = (\nu_1, \ldots, \nu_T)$ from repeated instances of a prompt about a random variable Y. To define optimal Bayesian aggregation, consider the benchmark in which the observer knows the prior μ_0 and the experiment E associated with the prompt. Write $\Pr$ for probability under this prior and experiment. The optimal belief aggregation is then given by the posterior prescribed by Bayes' rule as

$$\nu^T(h_T)(y) := \Pr(Y(\omega) = y \mid h_T) \quad \forall\, y \in \mathcal{Y}. \tag{6}$$

For any given h_T, we say that $\eta \in \Delta(\mathcal{Y})$ is a *Bayes-plausible aggregate after observing* h_T if there exist a full-support prior μ_0, a finite experiment E, and a true state $\omega \in \Omega$ such that h_T has positive probability conditional on ω and $\nu^T(h_T) = \eta$. For a full-support prior μ_0, we say that η is a μ_0-*Bayes-plausible aggregate after observing* h_T if the same conditions hold with the prior fixed at μ_0.

Theorem 3 (Bayes-Plausible Aggregation)**.** *Suppose that $|\Omega_y| \geq 2$ for every $y \in \mathcal{Y}$ and h_T contains at least two distinct reports.*[2] *For any full-support $\mu_0 \in \Delta(\Omega)$, $\eta \in \Delta(\mathcal{Y})$ is a μ_0-Bayes-plausible aggregate after observing h_T if and only if*

$$\operatorname{supp} \eta \subseteq \operatorname{supp} \nu_i \qquad \forall\, i = 1, \ldots, T. \tag{7}$$

Proof. See Appendix A.8. □

As in the previous sections, if the observer does not know the experiment, the only guaranteed inference comes from hard evidence: any answer value $y \in \mathcal{Y}$ excluded by some report must also be excluded by the aggregate. Also, because condition (7) does not depend on μ_0, it also characterizes Bayes-plausible aggregates when the prior is unknown.

Remark 6 (Aggregation Bottleneck)**.** This result contrasts with the familiar formula for aggregation with a known prior when the agent reports beliefs about the full state (see Lemma 3 in the appendix). In that setting, each report is a posterior μ over the complete state ω, repetitions are conditionally

[2] Lemma 4 in the Appendix characterizes the case of all reports being the same.

i.i.d. given that state, and the Bayesian aggregate is a prior-corrected sum of reported log odds: for any $\omega, \omega' \in \operatorname{supp} \nu^T(h_T)$,

$$\log \frac{\nu^T(h_T)(\omega)}{\nu^T(h_T)(\omega')} = \sum_{i=1}^{T} \log \frac{\nu_i(\omega)}{\nu_i(\omega')} - (T-1) \log \frac{\mu_0(\omega)}{\mu_0(\omega')}. \tag{8}$$

Importantly, the formula highlights that the experiment *need not be known* for optimal aggregation. Thus, it is the inability to communicate full posterior about the state of the world that creates a bottleneck for knowledge extraction.

The reason is that repetitions are independent conditional on the complete state ω, but generally not conditional on the coarser value $y = Y(\omega)$. The distribution is exchangeable in the repetitions, but the common latent state ω can correlate all reports after conditioning only on y. A full-state report identifies the state-by-state likelihood ratios needed in the product formula, whereas a coarse report identifies only their prior-weighted averages within each cell Ω_y. Those averages do not determine the likelihood of a history without knowing the experiment, which is why coarse-report aggregation is much less restrictive. ♢

Although Theorem 3 permits many aggregates, it excludes some natural rules. Arithmetic averaging is not Bayes-plausible whenever one report assigns probability zero to an answer and another assigns it positive probability. Affine log-odds aggregation respects this support restriction and is exactly optimal in a Gaussian model (see Appendix A.10).

Example 1 (ctd.)**.** We now show that even unlimited repetition can leave the Bayesian aggregate undetermined. Consider agent 2's report distribution in Figure 1. Split the previous state 0 into two and relabel the resulting space as $\Omega = \{0, 1, 2, 3\}$, with $Y(0) = Y(1) = 0$ and $Y(2) = Y(3) = 1$. Let the prior be $\mu_0 = (1, 99, 1, 19)/120$. Appendix B.2 constructs two fixed experiments with this prior and the same entire observed report distribution τ_2. In one rationalization state 2 is true and the Bayesian aggregate converges almost surely to 1; in the other state 0 is true and it converges almost surely to 0. Each experiment is fixed for all T, and both generate the same distribution over every finite report history. The empirical mean converges to $1/2$ in both cases, while the Bayesian aggregates converge to opposite answers.

By contrast, replacing the lower report by an exact zero forces the aggregate to assign zero probability to $Y = 1$ as soon as that report occurs. Thus, for purposes of aggregation, there is a substantive distinction between an arbitrarily small positive probability and zero. ♦

6 Future Directions

Our framework provides a theoretical basis for interpreting existing ML findings and developing more specialized models of reasoning. One direction is to extend our aggregation results by imposing additional structure on the internal experiments. Relevant applications include self-consistency, which selects the most frequent answer across reasoning attempts (Wang et al., 2023), and semantic uncertainty, which treats answers with the same meaning as one outcome when measuring uncertainty (Kuhn et al., 2023). Such models could clarify when agreement across sampled answers predicts correctness and how those answers should be combined.

A second direction is to ask what checking or revising an answer adds. Snell et al. (2025) study verifier-guided search and answer revision, finding that effective computation allocation depends on question difficulty. Our information order could help assess when these procedures leave the agent better informed. If each step retains all earlier signals, additional signals cannot reduce Blackwell informativeness, but informativeness can decrease if only a summary of earlier reasoning is retained.

A third direction is to elicit more information about the agent's internal experiment. Direct elicitation may be problematic, as the agent may not understand its own inner workings. Instead, simpler, indirect tests could build on existing tests of Bayesian updating (Falck et al., 2024; Kim et al., 2025). For example, one could ask for a belief and the agent's probabilities for possible evidence, then supply new evidence and ask for an updated belief.

Our results provide a theoretical foundation for these directions and highlight that rigorous conclusions require explicit assumptions about how agents reason and report.

DISCLOSURE OF AI USE

In this work, we used generative AI tools for literature search, copyediting, and proofreading. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Isaiah Andrews and Suproteem Sarkar. Dutch books for language models. arXiv:2609.02797, 2026. https://arxiv.org/abs/2609.02797.

Francis J. Anscombe and Robert J. Aumann. A definition of subjective probability. *The Annals of Mathematical Statistics*, 34(1):199–205, 1963.

Itai Arieli and Yakov Babichenko. A population's feasible posterior beliefs. *Journal of Economic Theory*, 215:105764, 2024.

Itai Arieli, Yakov Babichenko, and Rann Smorodinsky. Robust forecast aggregation. *Proceedings of the National Academy of Sciences*, 115(52):E12135–E12143, 2018.

Itai Arieli, Yakov Babichenko, and Rann Smorodinsky. Identifiable information structures. *Games and Economic Behavior*, 120:16–27, 2020.

Katherine Atwell, Pedram Heydari, Anthony Sicilia, and Malihe Alikhani. BASIL: Bayesian assessment of sycophancy in LLMs. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, pages 6613–6642, 2026.

Robert J. Aumann and Michael B. Maschler, with Richard E. Stearns. *Repeated Games with Incomplete Information*. MIT Press, Cambridge, MA, 1995.

Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

Siddharth Boppana, Annabel Ma, Max Loeffler, Raphael Sarfati, Eric Bigelow, Atticus Geiger, Owen Lewis, and Jack Merullo. Reasoning theater: Disentangling model beliefs from chain-of-thought. arXiv:2603.05488, 2026.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712, 2023.

Andrew Caplin, Daniel Martin, and Philip Marx. Modeling machine learning: A cognitive economic approach. *Journal of Economic Theory*, 224:105970, 2025.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. arXiv:2505.05410, 2025.

Chacha Chen, Matthew Jörke, Adam Goliński, Masha Fedzechkina, Guillermo Sapiro, Sinead Williamson, and Nicholas Foti. LLMs are not (consistently) Bayesian: Quantifying internal (in)consistencies of LLMs' probabilistic beliefs. arXiv:2605.06915, 2026.

Julian Coda-Forno, Marcel Binz, Jane X. Wang, and Eric Schulz. CogBench: A large language model walks into a psychology lab. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9076–9108, 2024.

Yuan Deng, Vahab Mirrokni, Renato Paes Leme, Hanrui Zhang, and Song Zuo. LLMs at the bargaining table. In *ICML 2024 Workshop on Agentic Markets*, 2024.

Laura Doval and Alex Smolin. Persuasion and welfare. *Journal of Political Economy*, 132(7):2451–2487, 2024.

Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 144–155, 2024.

Paul Dütting, Safwan Hossain, Tao Lin, Renato Paes Leme, Sai Srivatsa Ravindranath, Haifeng Xu, and Song Zuo. Information design with large language models. In *27th ACM Conference on Economics and Computation*, 2026. Full version: `https://arxiv.org/abs/2509.25565`.

Fabian Falck, Ziyu Wang, and Christopher C. Holmes. Is in-context learning in large language models Bayesian? A martingale perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12784–12805, 2024.

Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic collusion by large language models. In *27th ACM Conference on Economics and Computation*, 2026. Full version: `https://arxiv.org/abs/2404.00806`.

Alejandro Francetich and David M. Kreps. Bayesian inference does not lead you astray… on average. *Economics Letters*, 125(3):444–446, 2014.

Christian Genest. A characterization theorem for externally Bayesian groups. *The Annals of Statistics*, 12(3):1100–1105, 1984.

John C. Harsanyi and Reinhard Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press, Cambridge, MA, 1988.

Zhonghao He, Tianyi Qiu, Hirokazu Shirado, and Maarten Sap. Martingale score: An unsupervised metric for Bayesian rationality in LLM reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pages 152614–152642, 2025.

Sohaib Imran, Ihor Kendiukhov, Matthew Broerman, Aditya Thomas, Riccardo Campanella, Rob Lamb, and Peter M. Atkinson. Are LLM belief updates consistent with Bayes' theorem? In *ICML 2025 Workshop on Assessing World Models: Methods and Metrics for Evaluating Understanding*, 2025.

Joseph M. Kahn. A generative Bayesian model for aggregating experts' probabilities. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 301–308, 2004.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8849–8861, 2021.

Minsu Kim, Sangryul Kim, and James Thorne. From evidence to belief: A Bayesian epistemology approach to language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10578–10611, 2025.

Kevin B. Korb and Ann E. Nicholson. *Bayesian Artificial Intelligence*. Second edition, Chapman & Hall/CRC, Boca Raton, FL, 2010.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Dharshan Kumaran, Stephen M. Fleming, Larisa Markeeva, Joe Heyward, Andrea Banino, Mrinal Mathur, Razvan Pascanu, Simon Osindero, Benedetto De Martino, Petar Veličković, and Viorica Patraucean. Competing biases underlie overconfidence and underconfidence in LLMs. *Nature Machine Intelligence*, 8(4):614–627, 2026.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. arXiv:2307.13702, 2023.

Filip Matějka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298, 2015.

Mantas Mazeika, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Oliver Zhang, and Dan Hendrycks. Utility engineering: Analyzing and controlling emergent value systems in AIs. *Advances in Neural Information Processing Systems*, 38:140822–140868, 2026.

Pooya Molavi. The empirical content of Bayesianism. arXiv:2109.07007v4, 2024. First circulated in 2021 under the title *Tests of Bayesian Rationality*.

Pooya Molavi. The empirical content of Bayesian updating under misspecification. arXiv:2507.22775v2, 2026.

OpenAI. GPT-4 technical report. arXiv:2303.08774, 2023.

Vardhan Palod, Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. Performative thinking? The brittle correlation between cot length and problem complexity. arXiv:2509.07339, 2025.

Akshay Paruchuri, Jake Garrison, Shun Liao, John Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, and Daniel McDuff. What are the odds? Language models are capable of probabilistic reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11712–11733, 2024.

Leonard J. Savage. *The Foundations of Statistics*. John Wiley & Sons, New York, NY, 1954.

Timo Pierre Schrader, Lukas Lange, Simon Razniewski, and Annemarie Friedrich. QUITE: Quantifying uncertainty in natural language text in Bayesian reasoning scenarios. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2634–2652, 2024.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2024.

Eran Shmaya and Leeat Yariv. Experiments on decisions under uncertainty: A theoretical framework. *American Economic Review*, 106(7):1775–1801, 2016.

Christopher A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690, 2003.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *The Tenth International Conference on Learning Representations*, 2022.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024.

Jian-Qiao Zhu and Thomas L. Griffiths. Incoherent probability judgments in large language models. In *Proceedings of the 46th Annual Conference of the Cognitive Science Society*, pages 906–913, 2024.

A PROOFS

A.1 REDUCTION TO REPORTS ON A PARTITION

We can interpret an arbitrary prompt as requesting the value of a function of the agent's posterior, $f_p : \Delta(\Omega) \to R$. For a partition $\mathcal{Q}_p = \{Q_1, \dots, Q_K\}$ of Ω, define

$$A_p(\mu)(k) := \sum_{\omega \in Q_k} \mu(\omega), \qquad k = 1, \dots, K.$$

If f_p is constant on each fiber of A_p, it factors as $f_p = g_p \circ A_p$ for some $g_p : \Delta(\{1, \dots, K\}) \to R$. Such a partition always exists: the singleton partition makes A_p a bijection. Set $Y_p(\omega) = k$ on Q_k. Then $\nu = A_p(\mu)$ is the posterior of Y_p, and the original response is $r = g_p(\nu)$, with no change to the experiment.

A.2 KNOWN PRIOR AND FULL-STATE REPORTS

The next lemma characterizes the restrictions a known prior imposes on full-state reports. It also provides the implementation step in the proof of Theorem 1: we lift the observed reports to full-state posteriors and choose a common prior satisfying the lemma's conditions for every prompt.

Throughout this subsection, set $\mathcal{Y} = \Omega$ and let Y be the identity map. Fix a full-support prior μ_0 and a candidate true state ω. A finite-support distribution τ of full posteriors is (μ_0, ω)-*Bayes-plausible* if some finite experiment induces τ conditional on ω.

Whenever every $\nu \in \operatorname{supp} \tau$ satisfies $\nu(\omega) > 0$, define

$$H(\tau, \omega; z) := \sum_{\nu \in \operatorname{supp} \tau} \tau(\nu) \frac{\nu(z)}{\nu(\omega)}. \tag{9}$$

Lemma 1 (Known-Prior Bayes-Plausibility)**.** *Fix $\mu_0 \in \operatorname{int} \Delta(\Omega)$ and $\omega \in \Omega$. A finite-support full-posterior report distribution τ is (μ_0, ω)-Bayes-plausible if and only if*

$$\nu(\omega) > 0 \qquad \text{for every } \nu \in \operatorname{supp} \tau, \tag{10}$$

$$H(\tau, \omega; z) \le \frac{\mu_0(z)}{\mu_0(\omega)} \qquad \text{for every } z \in \Omega. \tag{11}$$

The inequality for $z = \omega$ always holds with equality.

Proof. Let $M := \operatorname{supp} \tau$. Suppose first that a finite experiment (S, π) induces τ at ω. Delete ex ante null signals, let $q(s) := \sum_x \mu_0(x)\pi(s \mid x)$, and let ν_s be the posterior induced by s. If $\pi(s \mid \omega) > 0$, then

$$\nu_s(\omega) = \frac{\mu_0(\omega)\pi(s \mid \omega)}{q(s)} > 0.$$

Every report in M is induced by at least one such signal, which proves (10).

For every $z \in \Omega$, Bayes' rule gives

$$H(\tau, \omega; z) = \sum_{s:\, \pi(s|\omega) > 0} \pi(s \mid \omega) \frac{\nu_s(z)}{\nu_s(\omega)} = \frac{\mu_0(z)}{\mu_0(\omega)} \sum_{s:\, \pi(s|\omega) > 0} \pi(s \mid z) \le \frac{\mu_0(z)}{\mu_0(\omega)}.$$

If several signals induce the same report, their probabilities at ω add to $\tau(\nu)$, so the first equality is valid. This proves necessity.

Conversely, suppose (10) and (11) hold. Introduce one signal s_ν for each $\nu \in M$ and set

$$\pi(s_\nu \mid z) := \frac{\mu_0(\omega)}{\mu_0(z)} \tau(\nu) \frac{\nu(z)}{\nu(\omega)}. \tag{12}$$

Then $\pi(s_\nu \mid \omega) = \tau(\nu)$ and, for every z,

$$\sum_{\nu \in M} \pi(s_\nu \mid z) = \frac{\mu_0(\omega)}{\mu_0(z)} H(\tau, \omega; z) \le 1.$$

Add a residual signal r with

$$\pi(r \mid z) := 1 - \sum_{\nu \in M} \pi(s_\nu \mid z),$$

omitting it if its likelihood is identically zero. Since $H(\tau, \omega; \omega) = 1$, the residual signal satisfies $\pi(r \mid \omega) = 0$.

The ex ante probability of s_ν is

$$q(s_\nu) = \sum_{z \in \Omega} \mu_0(z)\pi(s_\nu \mid z) = \frac{\mu_0(\omega)\tau(\nu)}{\nu(\omega)}.$$

Therefore

$$\Pr(z \mid s_\nu) = \frac{\mu_0(z)\pi(s_\nu \mid z)}{q(s_\nu)} = \nu(z).$$

Thus s_ν induces ν, occurs at ω with probability $\tau(\nu)$, and the residual signal never occurs at ω. The induced report distribution is exactly τ. □

The proof also shows that every rationalizing experiment satisfies

$$\frac{\mu_0(\omega)}{\mu_0(z)} H(\tau, \omega; z) = \sum_{s:\, \pi(s|\omega)>0} \pi(s \mid z) = 1 - \sum_{s:\, \pi(s|\omega)=0} \pi(s \mid z). \tag{13}$$

Hence (11) holds with equality for every z if and only if no signal possible at any state is impossible at ω. In particular, equality holds when the experiment has common signal support across states.

For a fixed candidate truth, the identified set of full-support priors is

$$\{\mu_0 \in \operatorname{int} \Delta(\Omega) : \mu_0(z) \geq \mu_0(\omega) H(\tau, \omega; z) \text{ for every } z \in \Omega\}. \tag{14}$$

Summing the inequalities gives

$$\mu_0(\omega) \leq \left(\sum_{z \in \Omega} H(\tau, \omega; z) \right)^{-1} = \left(\mathbb{E}_{\nu \sim \tau} \left[\frac{1}{\nu(\omega)} \right] \right)^{-1}. \tag{15}$$

Jensen's inequality then implies $\mathbb{E}_\tau[\nu(\omega)] \geq \mu_0(\omega)$. This is the *truth-drifting* property presented by Francetich and Kreps (2014) and Doval and Smolin (2024). The inverse-probability restriction is sharper than this mean restriction. In the binary case it is the full restriction. If $\Omega = \{0, 1\}$, $\mu_0(1) = q$, the true state is 1, and a report assigns probability r to state 1, Bayes-plausibility is equivalent to

$$\mathbb{E}_{r \sim \tau} \left[\frac{1 - r}{r} \right] \leq \frac{1 - q}{q}, \qquad \text{equivalently} \qquad \mathbb{E}_{r \sim \tau} \left[\frac{1}{r} \right] \leq \frac{1}{q}. \tag{16}$$

If the true state is not specified, τ is Bayes-plausible under μ_0 if and only if some $\omega \in \bigcap_{\nu \in \operatorname{supp} \tau} \operatorname{supp} \nu$ satisfies (11). If the prior is also unrestricted, choosing $\mu_0(\omega)$ sufficiently small makes all inequalities hold. Thus the lemma reduces to the single-prompt full-state case of Theorem 1 when both the prior and the true state may vary.

A.3 Proof of Theorem 1

Proof. Suppose first that the collection is Bayes-plausible, and let ω be the common true state. For every $p \in P$ and $\nu \in \operatorname{supp} \tau_p$, some signal inducing ν occurs at ω. Full support of the prior and Bayes' rule imply that the corresponding full posterior assigns positive probability to ω. Its Y_p-marginal therefore satisfies $\nu(Y_p(\omega)) > 0$. Hence ω belongs to the intersection in (3).

Conversely, suppose (3) holds and choose ω in that intersection. For each p, write $\Omega_p(y) := Y_p^{-1}(y)$ and lift each report $\nu \in \operatorname{supp} \tau_p$ by distributing its mass uniformly within each cell:

$$L_p(\nu)(\omega') := \frac{\nu(Y_p(\omega'))}{|\Omega_p(Y_p(\omega'))|}. \tag{17}$$

The posterior $L_p(\nu)$ has Y_p-marginal ν and assigns positive probability to ω. Let $\widehat{\tau}_p$ be its distribution when $\nu \sim \tau_p$, and write $H_p(z) := H(\widehat{\tau}_p, \omega; z)$. These odds moments are finite and satisfy $H_p(\omega) = 1$.

Choose a common prior by normalizing the positive weights

$$w(z) := \begin{cases} 1, & z = \omega, \\ 1 + \sum_{p \in P} H_p(z), & z \neq \omega, \end{cases} \qquad \mu_0(z) := \frac{w(z)}{\sum_{x \in \Omega} w(x)}.$$

Then, for every $p \in P$ and $z \in \Omega$,

$$H_p(z) \leq \frac{\mu_0(z)}{\mu_0(\omega)}. \tag{18}$$

By Lemma 1, each $\widehat{\tau}_p$ is induced at ω under this prior by a finite experiment. Reporting its Y_p-marginal induces τ_p, proving sufficiency, including when $|\Omega| = 1$. □

A.4 Proof of Theorem 2

For the intelligence-order results below, let

$$C := \bigcap_{\nu \in M_1 \cup M_2} \operatorname{supp} \nu.$$

Proof. Necessity. Delete ex ante null signals. Let ω be the true state, set $y := Y(\omega)$, and let $q_j(s_j)$ and $\nu_j(s_j)$ be the ex ante probability and posterior report associated with signal s_j. Bayes' rule gives

$$q_j(s_j)\nu_j(s_j)(z) = \sum_{\omega' \in \Omega_z} \mu_0(\omega')\pi_j(s_j \mid \omega') \qquad \text{for every } z \in \mathcal{Y}. \tag{19}$$

Every signal occurring at ω assigns positive probability to y, so $y \in C$.

Let G garble E_2 into E_1. Conditional on ω, draw s_2 and then s_1 from $G(\cdot \mid s_2)$. The induced distribution γ of $(\nu_2(s_2), \nu_1(s_1))$ is a coupling of τ_2 and τ_1. Moreover,

$$q_1(s_1)\nu_1(s_1) = \sum_{s_2 \in S_2} q_2(s_2)G(s_1 \mid s_2)\nu_2(s_2). \tag{20}$$

If (ν_2, ν_1) has positive γ-probability, some pair (s_2, s_1) with those reports satisfies $\pi_2(s_2 \mid \omega)G(s_1 \mid s_2) > 0$. If $\nu_1(z) = 0$, equation (20) and nonnegativity force $\nu_2(z) = 0$. Hence $\operatorname{supp} \nu_2 \subseteq \operatorname{supp} \nu_1$ on the support of γ.

Sufficiency. Fix $y \in C$, a coupling γ satisfying (ii), and a state $\omega \in \Omega_y$. We first give an implementation step.

Consider finitely many abstract signals s. Assign each signal an ex ante probability $q(s) > 0$, a probability $a(s) \geq 0$ at ω, and a report $v_s \in \Delta(\mathcal{Y})$. Suppose

$$\sum_s q(s) = \sum_s a(s) = 1, \qquad m := \sum_s q(s)v_s \in \operatorname{int} \Delta(\mathcal{Y}),$$

and there is $x \in (0, m(y))$ such that

$$q(s)v_s(y) \geq xa(s) \qquad \text{for every } s.$$

Choose a full-support prior with cell masses $\mu_0(\Omega_z) = m(z)$, assign mass x to ω, and distribute the remaining mass $m(y) - x$ strictly positively over $\Omega_y \setminus \{\omega\}$. This is possible because $|\Omega_y| \geq 2$. Define

$$\pi(s \mid \omega') := \begin{cases} a(s), & \omega' = \omega, \\ q(s)v_s(z)/m(z), & \omega' \in \Omega_z,\ z \neq y, \\ \big(q(s)v_s(y) - xa(s)\big)/(m(y) - x), & \omega' \in \Omega_y \setminus \{\omega\}. \end{cases}$$

The likelihoods are nonnegative and sum to one at every state. Signal s has ex ante probability $q(s)$ and induces report v_s.

We now construct the abstract signals for E_2. Fix $c := 1/2$. By the support condition, one can choose $\varepsilon \in (0, c)$ so small that, for every $\nu_1 \in M_1$,

$$w(\nu_1) := c\tau_1(\nu_1)\nu_1 - \varepsilon \sum_{\nu_2 \in M_2} \gamma(\nu_2, \nu_1)\nu_2$$

is nonnegative and strictly positive on $\operatorname{supp} \nu_1$. A nonzero vector $w \in \mathbb{R}^{\mathcal{Y}}_+$ represents a signal with ex ante probability $\|w\|_1$ and report $w/\|w\|_1$. Since γ has second marginal τ_1,

$$\|w(\nu_1)\|_1 = (c - \varepsilon)\tau_1(\nu_1).$$

For every pair with $\gamma(\nu_2, \nu_1) > 0$, create a signal with report ν_2, ex ante probability $\varepsilon\gamma(\nu_2, \nu_1)$, and probability $\gamma(\nu_2, \nu_1)$ at ω. Garble it to a destination indexed by ν_1. For each ν_1, add the signal represented by $w(\nu_1)$, give it probability zero at ω, and garble it to the same destination. Finally, add a signal with an interior report, ex ante probability $1 - c$, probability zero at ω, and a separate destination.

The ex ante probabilities sum to

$$\varepsilon + (c - \varepsilon) + (1 - c) = 1,$$

and the probabilities at ω sum to one. The mean report is interior because the last signal has positive ex ante probability and an interior report. Choose

$$0 < x < \varepsilon \min_{\nu_2 \in M_2} \nu_2(y).$$

For a signal indexed by (ν_2, ν_1), the implementation inequality is

$$\varepsilon\gamma(\nu_2, \nu_1)\nu_2(y) \geq x\gamma(\nu_2, \nu_1).$$

It is automatic for all signals with probability zero at ω. Moreover, the pair signals alone give the mean report more than x at y. The implementation step therefore produces E_2.

At the destination indexed by ν_1, the prior-weighted report vector is

$$\varepsilon \sum_{\nu_2 \in M_2} \gamma(\nu_2, \nu_1)\nu_2 + w(\nu_1) = c\tau_1(\nu_1)\nu_1,$$

while its probability at ω is $\sum_{\nu_2} \gamma(\nu_2, \nu_1) = \tau_1(\nu_1)$. Thus the garbled report is ν_1. The ungarbled report distribution at ω is τ_2, the garbled report distribution is τ_1, and E_2 Blackwell dominates E_1. □

A.5 Proof of the Different-Priors Remark

Modify the definition of plausible Blackwell dominance by allowing agent i to have its own full-support prior μ_i while retaining a common true state. If $|\Omega_y| \geq 2$ for every $y \in \mathcal{Y}$, conditions (i) and (ii) continue to characterize plausible Blackwell dominance.

Proof. For each ex ante non-null signal s_i, write $\nu_i(s_i)$ for agent i's report. Bayes' rule and full support of μ_i give

$$z \in \operatorname{supp} \nu_i(s_i) \quad \Longleftrightarrow \quad \pi_i(s_i \mid \omega') > 0 \text{ for some } \omega' \in \Omega_z. \tag{21}$$

Thus posterior support depends only on the experiment's likelihoods, not on the positive weights in the prior.

For necessity, let ω be the common true state and let G garble E_2 into E_1. As in the necessity proof of Theorem 2, $Y(\omega) \in C$, and drawing s_2 at ω followed by s_1 from G induces a coupling γ of the report distributions. This construction does not require a common prior. Every pair in the support of γ is induced by a signal pair with $\pi_2(s_2 \mid \omega)G(s_1 \mid s_2) > 0$. For any $z \in \operatorname{supp} \nu_2(s_2)$, (21) gives an $\omega' \in \Omega_z$ such that $\pi_2(s_2 \mid \omega') > 0$. Therefore

$$\pi_1(s_1 \mid \omega') = \sum_{t \in S_2} \pi_2(t \mid \omega')G(s_1 \mid t) \geq \pi_2(s_2 \mid \omega')G(s_1 \mid s_2) > 0.$$

Another application of (21) yields $z \in \operatorname{supp} \nu_1(s_1)$, proving the required support inclusion.

Conversely, if the two conditions hold, Theorem 2 supplies a rationalization with a common full-support prior. This is also admissible when the agents are allowed, but not required, to have different priors. □

A.6 Proof of Corollary 1

Proof. Let ν^0 and ν^1 be the two extreme beliefs. Every other report has support $\mathcal{Y}$, so the common-truth condition in Theorem 2 holds exactly when $M_1 \cup M_2$ contains at most one of ν^0 and ν^1. Under a support-respecting coupling, a less informative extreme report ν^k can only be paired with the same more informative report, implying $\tau_2(\nu^k) \geq \tau_1(\nu^k)$. Conversely, if these inequalities hold, match the mass $\tau_1(\nu^k)$ at each extreme belief and couple all residual mass to the nonextreme reports under τ_1; every such pair satisfies support inclusion. The result follows from Theorem 2. □

Lemma 2 (Intelligence Order for $|\Omega_y| = 1$)**.** *Fix $y \in \mathcal{Y}$ such that $\Omega_y = \{\omega\}$. There exist a full-support prior and experiments E_1, E_2 such that E_2 Blackwell dominates E_1 and the report distribution under E_i, conditional on ω, is τ_i for $i \in \{1, 2\}$ if and only if $y \in C$ and there is a coupling γ of τ_2 and τ_1 satisfying, coordinatewise,*

$$\sum_{\nu_2 \in M_2} \gamma(\nu_2, \nu_1) \frac{\nu_2}{\nu_2(y)} \leq \tau_1(\nu_1) \frac{\nu_1}{\nu_1(y)} \qquad \text{for every } \nu_1 \in M_1. \tag{22}$$

This inequality implies $\operatorname{supp} \nu_2 \subseteq \operatorname{supp} \nu_1$ *whenever* $\gamma(\nu_2, \nu_1) > 0$.

Proof. The necessity argument for Theorem 2 gives $y \in C$ and a signal coupling. Let $t := \mu_0(\omega)$. Since $\Omega_y = \{\omega\}$, (19) gives

$$q_j(s_j)\nu_j(s_j)(y) = t\pi_j(s_j \mid \omega).$$

Fix $\nu_1 \in M_1$, sum (20) over all signals inducing ν_1, and retain only signal pairs that occur at ω. All omitted terms are nonnegative, so

$$t\tau_1(\nu_1) \frac{\nu_1}{\nu_1(y)} \geq t \sum_{\nu_2 \in M_2} \gamma(\nu_2, \nu_1) \frac{\nu_2}{\nu_2(y)}.$$

Canceling $t > 0$ proves (22). If $\nu_1(z) = 0$, this condition and nonnegativity imply $\nu_2(z) = 0$ for every ν_2 paired with ν_1, proving the stated support implication.

For sufficiency, the case $|\mathcal{Y}| = 1$ is immediate. Suppose $|\mathcal{Y}| \geq 2$ and choose $x > 0$ such that

$$x \sum_{\nu_1 \in M_1} \frac{\tau_1(\nu_1)}{\nu_1(y)} < 1.$$

For every pair with $\gamma(\nu_2, \nu_1) > 0$, create a signal with report ν_2, probability $\gamma(\nu_2, \nu_1)$ at ω, and ex ante probability

$$q(\nu_2, \nu_1) := x \frac{\gamma(\nu_2, \nu_1)}{\nu_2(y)}.$$

Garble it to a destination indexed by ν_1. Define the residual vector

$$w(\nu_1) := x \left(\tau_1(\nu_1) \frac{\nu_1}{\nu_1(y)} - \sum_{\nu_2 \in M_2} \gamma(\nu_2, \nu_1) \frac{\nu_2}{\nu_2(y)} \right).$$

It is nonnegative by (22), and its y-coordinate is zero. If $w(\nu_1) \neq 0$, add the signal represented by this vector to the same destination and give it probability zero at ω. Summing the pair masses and the ℓ_1-norms of the residual vectors gives total ex ante probability

$$x \sum_{\nu_1 \in M_1} \frac{\tau_1(\nu_1)}{\nu_1(y)} < 1.$$

Use the remaining probability for a signal that has probability zero at ω, assigns zero probability to y, assigns positive probability to every other value, and has a separate destination.

Let $q(s)$, $a(s)$, and v_s denote the ex ante probability, probability at ω, and report of each signal. Their ex ante probabilities and probabilities at ω both sum to one. Their mean report $m := \sum_s q(s)v_s$ is interior, $m(y) = x$, and

$$q(s)v_s(y) = xa(s) \qquad \text{for every } s.$$

Give each cell Ω_z prior mass $m(z)$, distributing it strictly positively within the cell. In particular, $\mu_0(\omega) = x$. Define

$$\pi_2(s \mid \omega') := \begin{cases} a(s), & \omega' = \omega, \\ q(s)v_s(z)/m(z), & \omega' \in \Omega_z,\ z \neq y. \end{cases}$$

These likelihoods sum to one at every state. Since $q(s)v_s(y) = xa(s)$, signal s has ex ante probability $q(s)$ and induces report v_s.

At the destination indexed by ν_1, the prior-weighted report vector is $x\tau_1(\nu_1)\nu_1/\nu_1(y)$ and the probability at ω is $\tau_1(\nu_1)$. Hence its report is ν_1. The ungarbled and garbled report distributions at ω are τ_2 and τ_1. □

A.7 Full-Posterior Aggregation

Lemma 3 (Full-Posterior Aggregation)**.** *Suppose that the prompt asks the agent to report its full posterior over the state of the world, so that each $\nu_i \in \Delta(\Omega)$. Then, for every history h_T that occurs with positive probability,*

$$\nu^T(h_T)(\omega) = \frac{\mu_0(\omega)^{1-T}\prod_{i=1}^T \nu_i(\omega)}{\sum_{\omega'\in\Omega}\mu_0(\omega')^{1-T}\prod_{i=1}^T \nu_i(\omega')} \qquad \textit{for every } \omega \in \Omega. \tag{23}$$

In particular, for any $\omega, \omega' \in \operatorname{supp}\nu^T(h_T)$,

$$\log\frac{\nu^T(h_T)(\omega)}{\nu^T(h_T)(\omega')} = \sum_{i=1}^T \log\frac{\nu_i(\omega)}{\nu_i(\omega')} - (T-1)\log\frac{\mu_0(\omega)}{\mu_0(\omega')}. \tag{24}$$

Proof. Let $q(\nu)$ be the ex ante probability of report ν. Since the report is the posterior itself, Bayes' rule gives

$$\Pr(\text{report} = \nu \mid \omega) = q(\nu)\frac{\nu(\omega)}{\mu_0(\omega)}. \tag{25}$$

Conditional independence therefore implies

$$\Pr(\omega \mid h_T) \propto \mu_0(\omega)\prod_{i=1}^T q(\nu_i)\frac{\nu_i(\omega)}{\mu_0(\omega)} \propto \mu_0(\omega)^{1-T}\prod_{i=1}^T \nu_i(\omega).$$

Normalizing proves (23); taking log odds proves (24). □

A.8 Proof of Theorem 3

Proof. Let $M := \{\nu_i : i = 1, \dots, T\}$ be the set of distinct observed reports, and let $n(\nu)$ be the multiplicity of $\nu \in M$. For each $\nu \in M$, let

$$\mu^\nu(\omega) := \Pr(\omega \mid \text{report} = \nu)$$

be its full-state posterior lift. Its Y-marginal is the report:

$$\sum_{\omega\in\Omega_y}\mu^\nu(\omega) = \nu(y) \qquad \text{for every } \nu \in M \text{ and } y \in \mathcal{Y}. \tag{26}$$

If q_ν is the ex ante probability of report ν, then

$$\Pr(\text{report} = \nu \mid \omega) = q_\nu \frac{\mu^\nu(\omega)}{\mu_0(\omega)}.$$

Conditional independence gives

$$\Pr(\omega \mid h_T) \propto \mu_0(\omega)^{1-T} \prod_{\nu \in M} \mu^\nu(\omega)^{n(\nu)}. \tag{27}$$

Thus, if

$$W_y := \sum_{\omega \in \Omega_y} \mu_0(\omega)^{1-T} \prod_{\nu \in M} \mu^\nu(\omega)^{n(\nu)}, \tag{28}$$

then the aggregate assigns probability $W_y / \sum_z W_z$ to y. If this probability is positive, some $\omega \in \Omega_y$ satisfies $\mu^\nu(\omega) > 0$ for every $\nu \in M$. By (26), every observed report then assigns positive probability to y. This proves necessity.

We next record a fixed-prior implementation step. Consider any finite collection of full-state posteriors $(\mu^\nu)_{\nu \in M}$ satisfying (26). Choose numbers $p_\nu > 0$ so small that

$$\sum_{\nu \in M} p_\nu < 1, \qquad \sum_{\nu \in M} p_\nu \mu^\nu(\omega) < \mu_0(\omega) \quad \text{for every } \omega.$$

Set

$$p_r := 1 - \sum_{\nu \in M} p_\nu, \qquad \mu^r := \frac{\mu_0 - \sum_{\nu \in M} p_\nu \mu^\nu}{p_r}.$$

Then μ^r has full support. The theorem's hypothesis of two distinct reports implies $|\mathcal{Y}| \geq 2$. Choose states $a, b \in \Omega$ with $Y(a) \neq Y(b)$, and let $d := \delta(a) - \delta(b)$. For all sufficiently small $\varepsilon > 0$, the distributions

$$\mu^+ := \mu^r + \varepsilon d, \qquad \mu^- := \mu^r - \varepsilon d$$

have full support. By excluding finitely many values of ε, their Y-marginals can also be chosen outside the finite set M. Define

$$\pi(s_\nu \mid \omega) := p_\nu \frac{\mu^\nu(\omega)}{\mu_0(\omega)}, \qquad \pi(r^\pm \mid \omega) := \frac{p_r}{2} \frac{\mu^\pm(\omega)}{\mu_0(\omega)}. \tag{29}$$

These likelihoods sum to one at every state because

$$\sum_{\nu \in M} p_\nu \mu^\nu + \frac{p_r}{2}(\mu^+ + \mu^-) = \mu_0.$$

Bayes' rule gives posterior μ^ν after s_ν, while the residual signals produce reports outside M. Thus any such collection of lifts is implementable under the fixed prior, with no additional signal producing an observed report.

For sufficiency, let $D := \operatorname{supp} \eta \subseteq \bigcap_{\nu \in M} \operatorname{supp} \nu$. Choose a report $\beta \in M$ and distinct states $a_y, b_y \in \Omega_y$ for every y. For parameters $t_y \in [0, \beta(y)]$, define the lifts in every cell by

$$\mu^\nu(a_y) := \nu(y) \qquad \text{for } \nu \neq \beta, \qquad \mu^\beta(a_y) := t_y, \qquad \mu^\beta(b_y) := \beta(y) - t_y.$$

Set every unspecified coordinate to zero. These lifts satisfy (26). Since M contains a report other than β, only a_y can contribute to the cell weight W_y.

For $y \in D$, define

$$A_y := \mu_0(a_y)^{1-T} \prod_{\nu \in M \setminus \{\beta\}} \nu(y)^{n(\nu)} > 0.$$

Choose $c > 0$ small enough that

$$t_y := \left(\frac{c\eta(y)}{A_y}\right)^{1/n(\beta)} < \beta(y) \qquad \text{for every } y \in D.$$

Then (28) gives $W_y = A_y t_y^{n(\beta)} = c\eta(y)$ on D. Set $t_y = 0$ off D, so that $W_y = 0$ there. The resulting aggregate is η.

The fixed-prior implementation step realizes these lifts. For any $y \in D$, every observed report has positive conditional probability at the state a_y, so the history occurs with positive probability there. This proves sufficiency. □

Lemma 4 (Bayes-Plausible Aggregation of Identical Reports)**.** *Suppose $|\Omega_y| \geq 2$ for every $y \in \mathcal{Y}$ and $h_T = (\nu, \ldots, \nu)$. For $T = 1$, the unique Bayes-plausible aggregate is $\eta = \nu$, also under any fixed full-support prior. For $T \geq 2$, an aggregate η is Bayes-plausible if and only if*

$$\operatorname{supp}\eta = \operatorname{supp}\nu. \tag{30}$$

For a fixed full-support prior μ_0, define

$$L_y := \frac{\nu(y)^T}{\mu_0(\Omega_y)^{T-1}}, \qquad U_y := \frac{\nu(y)^T}{\left(\min_{\omega\in\Omega_y}\mu_0(\omega)\right)^{T-1}} \quad (y \in \operatorname{supp}\nu). \tag{31}$$

Then η is μ_0-Bayes-plausible if and only if (30) *holds and some $c > 0$ satisfies*

$$L_y \leq c\eta(y) \leq U_y \quad \textit{for all } y \in \operatorname{supp}\nu, \tag{32}$$

or, equivalently,

$$\max_{y\in\operatorname{supp}\nu} \frac{L_y}{\eta(y)} \leq \min_{y\in\operatorname{supp}\nu} \frac{U_y}{\eta(y)}. \tag{33}$$

A.9 Proof of Lemma 4

Proof. The case $|\mathcal{Y}| = 1$ is immediate. Otherwise, the construction in (29) implements any full-state lift x of ν under μ_0. Choosing any state in $\operatorname{supp} x$ as the true state makes the repeated report occur with positive probability. By (27), its unnormalized aggregate weights are

$$W_y(x) = \sum_{\omega\in\Omega_y} \frac{x(\omega)^T}{\mu_0(\omega)^{T-1}}, \qquad \sum_{\omega\in\Omega_y} x(\omega) = \nu(y). \tag{34}$$

For $T = 1$, these weights equal $\nu(y)$. For $T \geq 2$, they are positive exactly on $\operatorname{supp}\nu$. Within each cell, Hölder's inequality gives the minimum L_y, attained at $x(\omega) = \nu(y)\mu_0(\omega)/\mu_0(\Omega_y)$. Convexity gives the maximum U_y, attained by concentrating mass on a state minimizing μ_0. Continuity on the connected cell simplex gives every intermediate value. Since cells can be treated independently, their normalized weights equal η exactly when (32) holds, equivalently (33).

For unrestricted sufficiency when $T \geq 2$, assume (30). Choose $a_y \in \Omega_y$ for each $y \in \operatorname{supp}\nu$, set $x(a_y) = \nu(y)$ and $x = 0$ elsewhere, and choose a full-support prior with

$$\mu_0(a_y) = \kappa\left(\frac{\nu(y)^T}{\eta(y)}\right)^{1/(T-1)}.$$

Take $\kappa > 0$ small enough that these masses sum to less than one and distribute the remainder strictly positively over all other states, which exist because $|\Omega_y| \geq 2$. Then $W_y(x) = \kappa^{1-T}\eta(y)$, proving sufficiency. □

A.10 Affine Log-Odds Aggregation

Fix a history $h_T = (\nu_1, \ldots, \nu_T)$ of reports about Y. Every Bayes-plausible aggregate must be supported on $D := \bigcap_{i=1}^T \operatorname{supp}\nu_i$. If $D = \varnothing$, no Bayes-plausible aggregate exists. Otherwise, fix

a base value $y_0 \in D$ and enumerate the remaining values in D as $\{y_1, \ldots, y_K\}$. Define the report log-odds vectors and their average by

$$l_i(k) := \log \frac{\nu_i(y_k)}{\nu_i(y_0)}, \quad \bar{l} := \frac{1}{T}\sum_{i=1}^{T} l_i, \quad m(k) := \log \frac{\nu(y_k)}{\nu(y_0)}, \quad i = 1, \ldots, T, \quad k = 1, \ldots, K.$$

For parameters $c \in \mathbb{R}^K$ and $C \in \mathbb{R}^{K \times K}$, affine log-odds aggregation sets

$$m = c + C\bar{l}. \tag{35}$$

The log-odds vector m uniquely determines the aggregate ν with support exactly D. Under the hypotheses of Theorem 3, the resulting ν is μ_0-Bayes-plausible for every full-support prior μ_0, and therefore Bayes-plausible. In (35), roughly, c captures the prior knowledge and C captures the learning patterns across $y \in \mathcal{Y}$. The choice $c = 0$ and $C = I_K$ gives ordinary average log odds, the equal-weight logarithmic opinion pool. This pool is externally Bayesian: pooling commutes with updating each input distribution by the same likelihood (Genest, 1984). The choice $c = (1 - T)l_0$ and $C = TI_K$ corresponds to the standard Bayesian aggregation rule with a known prior log odds l_0 and conditionally independent reports. More generally, affine log-odds aggregation is Bayes-optimal in the Gaussian intelligence model presented in Appendix A.11. In its homogenous case, $C = \lambda I_K$ and $c = (1 - \lambda)l_0$, where $\lambda \in [1, T]$ captures the correlation across reports.

A.11 Gaussian Intelligence

Conditional independence in the model is imposed given the full state ω, not generally given the coarser value $y = Y(\omega)$. Repeated reports may therefore remain correlated conditional on y because they share components of ω not revealed by y. The following Gaussian specification makes this dependence explicit. Related generative Bayesian aggregators for independent and exchangeable experts are developed by Kahn (2004).

Allow the full state to have a continuous component and signals to have continuous realizations. Let $J := |\mathcal{Y}|$ and, without loss, $\mathcal{Y} = \{1, \ldots, J\}$; set $K := J - 1$. For any full-support posterior $\nu \in \Delta(\mathcal{Y})$, write its log odds relative to J as

$$l(k) := \log \frac{\nu(k)}{\nu(J)}, \qquad k = 1, \ldots, K,$$

and let $l_0(k) := \log(\Pr(y = k)/\Pr(y = J))$ be the prior log odds. Let $e_1, \ldots, e_K$ be the standard basis of $\mathbb{R}^K$ and set $e_J := 0$. Fix symmetric matrices $G, \Sigma \in \mathbb{R}^{K \times K}$ with $G \succ 0$ and $0 \preceq \Sigma \preceq G$.

Let the full state be (y, u), where $u \sim N(0, \Sigma)$ independently of y. Conditional on (y, u), repetition i produces

$$s_i = l_0 - \frac{1}{2}\operatorname{diag}(G) + Ge_y + u + \varepsilon_i, \qquad i = 1, \ldots, T, \tag{36}$$

where $\varepsilon_i \sim N(0, G - \Sigma)$ independently across i and independently of (y, u). Thus the signals are i.i.d. conditional on the full state and exchangeable conditional on y.

To verify calibration, let

$$m_J := l_0 - \frac{1}{2}\operatorname{diag}(G), \qquad m_k := m_J + Ge_k.$$

Then $s_i \mid y = k \sim N(m_k, G)$. If f_k denotes its density, then, for $k = 1, \ldots, K$,

$$\log \frac{f_k(s)}{f_J(s)} = (m_k - m_J)^\top G^{-1}\left(s - \frac{m_k + m_J}{2}\right) = s(k) - l_0(k).$$

Adding prior log odds gives

$$\log \frac{\Pr(y = k \mid s_i)}{\Pr(y = J \mid s_i)} = s_i(k).$$

Hence the signal itself is the posterior log-odds vector, so write $l_i := s_i$.

For aggregation, define

$$\bar{l}_T := \frac{1}{T}\sum_{i=1}^{T} l_i, \qquad V_T := \Sigma + \frac{G-\Sigma}{T}.$$

Since $V_T = G/T + (1-1/T)\Sigma \succ 0$, this covariance is invertible even when Σ or $G-\Sigma$ is singular. Conditional on $y=k$, $\bar{l}_T \sim N(m_k, V_T)$. Moreover,

$$\mathrm{Cov}(\bar{l}_T, l_i - \bar{l}_T \mid y) = 0.$$

Conditional on y, the average and the centered residuals are jointly Gaussian and therefore independent. The distribution of the residuals does not depend on y, so $\bar{l}_T$ is sufficient for y.

Proposition 1 (Gaussian aggregation)**.** *In the Gaussian intelligence model, there exist a vector $c_T(l_0, \Sigma, G) \in \mathbb{R}^K$ and a matrix $C_T(\Sigma, G) \in \mathbb{R}^{K\times K}$ such that the posterior log odds after observing $l_1, \ldots, l_T$ are*

$$l^T = c_T(l_0, \Sigma, G) + C_T(\Sigma, G)\frac{1}{T}\sum_{i=1}^{T} l_i. \tag{37}$$

Proof. Let g_k be the conditional density of $\bar{l}_T$ given $y=k$. Since $m_k - m_J = Ge_k$,

$$\begin{aligned}\log\frac{g_k(\bar{l}_T)}{g_J(\bar{l}_T)} &= (m_k - m_J)^\top V_T^{-1}\left(\bar{l}_T - \frac{m_k+m_J}{2}\right)\\ &= e_k^\top G V_T^{-1}\left(\bar{l}_T - l_0 + \frac{1}{2}\operatorname{diag}(G)\right) - \frac{1}{2}e_k^\top G V_T^{-1} G e_k.\end{aligned}$$

Adding the prior log odds and stacking the coordinates gives

$$l^T = l_0 + GV_T^{-1}\left(\bar{l}_T - l_0 + \frac{1}{2}\operatorname{diag}(G)\right) - \frac{1}{2}\operatorname{diag}(GV_T^{-1}G).$$

Thus (37) holds with

$$\begin{aligned}C_T(\Sigma, G) &:= GV_T^{-1},\\ c_T(l_0, \Sigma, G) &:= l_0 - GV_T^{-1}l_0 + \frac{1}{2}GV_T^{-1}\operatorname{diag}(G) - \frac{1}{2}\operatorname{diag}(GV_T^{-1}G).\end{aligned}$$

□

Corollary 2 (Homogenous Gaussian aggregation)**.** *Suppose that $\Sigma = \rho G$ for some $\rho \in [0,1]$. Then*

$$l^T = (1-\lambda_T)l_0 + \lambda_T \bar{l}_T, \qquad \lambda_T := \frac{T}{1+(T-1)\rho}.$$

Conversely, if $T \geq 2$, then $C_T(\Sigma, G)$ is a scalar multiple of the identity if and only if Σ is proportional to G.

Proof. If $\Sigma = \rho G$, then

$$V_T = \frac{1+(T-1)\rho}{T}G,$$

so $C_T = GV_T^{-1} = \lambda_T I_K$. Substitution into the expression for c_T gives $c_T = (1-\lambda_T)l_0$.

Conversely, suppose $T \geq 2$ and $GV_T^{-1} = \lambda I_K$. Since G and V_T are positive definite, $\lambda > 0$ and $V_T = \lambda^{-1}G$. Using $TV_T = G + (T-1)\Sigma$ shows that Σ is proportional to G. □

If $\Sigma = 0$, the formula reduces to $l^T = T\bar{l}_T - (T-1)l_0$, the expression for conditionally independent reports. If $\Sigma = G$, all reports coincide and $l^T = \bar{l}_T = l_1$.

B RUNNING EXAMPLE

Throughout this example, r denotes the reported probability of $Y = 1$, and the two report distributions at the realized state are

$$\tau_1 = \tfrac{1}{2}\delta_{2/5} + \tfrac{1}{2}\delta_{3/5}, \qquad \tau_2 = \tfrac{1}{2}\delta_{1/10} + \tfrac{1}{2}\delta_{9/10}.$$

B.1 OPPOSITE INTELLIGENCE ORDERS

On a two-state space, neither Blackwell order is possible for these distributions, for any common full-support prior. At realized state 1, Lemma 2 implies that $E_2 \succeq_B E_1$ requires

$$\mathbb{E}_{\tau_2}\left[\frac{1-r}{r}\right] \le \mathbb{E}_{\tau_1}\left[\frac{1-r}{r}\right],$$

but the two sides are $41/9$ and $13/12$. Conversely, every ex ante possible full posterior under an experiment inducing τ_1 at state 1 lies in $\{0, 2/5, 3/5\}$: any signal absent at state 1 has posterior 0. A garbling cannot produce the report $9/10$, ruling out $E_1 \succeq_B E_2$. Symmetry gives both impossibilities when the realized state is 0.

Now let $\Omega = \{0, 1, 2\}$, with $Y(0) = 0$ and $Y(1) = Y(2) = 1$, realized state 1, and common full-state prior

$$\mu_0 = \tfrac{1}{120}(100, 1, 19).$$

With rows ordered as $(0, 1, 2)$, consider the conditional signal matrices

$$\pi_2 = \begin{pmatrix} 9/200 & 1/200 & 3/20 & 4/5 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 4/19 & 15/19 & 0 \end{pmatrix}, \qquad \pi_1 = \begin{pmatrix} 3/25 & 2/25 & 4/5 \\ 1/2 & 1/2 & 0 \\ 15/38 & 23/38 & 0 \end{pmatrix}. \tag{38}$$

The experiment E_2 from (5) induces reports $(1/10, 9/10, 1/2, 0)$. Send its first two signals to separate destinations, split its third signal equally between those destinations, and send its fourth signal to a residual destination. This garbling gives E_1, whose reports are $(2/5, 3/5, 0)$. At state 1, each experiment generates its first two reports with probability $1/2$, giving exactly the plotted distributions. The prior over Y remains $(5/6, 1/6)$. The order is strict: the ex ante squared-error Bayes risks for predicting Y are $7/100$ under E_2 and $8/100$ under E_1.

The reverse strict order is possible at the same full-state prior and realized state. Consider experiments $\widehat{E}_1, \widehat{E}_2$ with conditional signal matrices

$$\widehat{\pi}_1 = \begin{pmatrix} 3/100 & 1/50 & 19/20 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 3/38 & 5/38 & 0 & 15/19 \end{pmatrix}, \qquad \widehat{\pi}_2 = \begin{pmatrix} 9/50 & 1/50 & 4/5 \\ 1/2 & 1/2 & 0 \\ 3/38 & 35/38 & 0 \end{pmatrix}. \tag{39}$$

The reports under $\widehat{E}_1$ are $(2/5, 3/5, 0, 1)$. Send its first two signals to separate destinations; send its third signal to the first destination with probability $3/19$ and to a residual destination otherwise; send its fourth signal to the second destination. The resulting experiment is $\widehat{E}_2$, with reports $(1/10, 9/10, 0)$. The realized report distributions are again (τ_1, τ_2), while squared-error Bayes risks are now $2/100$ for $\widehat{E}_1$ and $3/100$ for $\widehat{E}_2$.

For the endpoint comparison, keep the same three-state space and allow the common prior and experiments to vary. Replacing τ_2 by $\frac{1}{2}\delta_{1/10} + \frac{1}{2}\delta_1$ leaves $Y = 1$ as the common possible answer, with two states in its cell. By the proof of Corollary 1, this distribution plausibly dominates τ_1, whereas the reverse order is impossible. Replacing its other atom by 0 leaves no common possible answer and violates Theorem 1.

B.2 AGGREGATION WITH ARBITRARILY MANY REPORTS

To study unlimited repetition, we construct two fixed experiments with the same prior and observed report distribution but opposite limiting aggregates.

Let $\Omega = \{0, 1, 2, 3\}$, with $Y(0) = Y(1) = 0$ and $Y(2) = Y(3) = 1$, and fix

$$\mu_0 = \tfrac{1}{120}(1, 99, 1, 19).$$

Consider experiments E^+, E^- with conditional signal matrices (rows ordered as $(0, 1, 2, 3)$)

$$\pi^+ = \begin{pmatrix} 0 & 0 & 1 \\ 1/22 & 1/198 & 94/99 \\ 1/2 & 1/2 & 0 \\ 0 & 4/19 & 15/19 \end{pmatrix}, \qquad \pi^- = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 4/99 & 0 & 95/99 \\ 0 & 0 & 1 \\ 1/38 & 9/38 & 14/19 \end{pmatrix}. \tag{40}$$

At the stated prior, both experiments have ex ante signal probabilities $(1/24, 1/24, 11/12)$ and induce reports $(1/10, 9/10, 3/22)$. The realized state is 2 under E^+ and 0 under E^-. Each generates the entire report distribution τ_2.

Let $n_L, n_H \geq 1$ count the reports $1/10, 9/10$, with $T = n_L + n_H$. The two aggregates are

$$\eta_T^+ = \frac{1}{1 + 99\,11^{-n_L}99^{-n_H}}, \qquad \eta_T^- = \frac{9^{n_H}}{19^{T-1} + 9^{n_H}}. \tag{41}$$

Under either realized distribution, both reports eventually appear almost surely. For every subsequent T,

$$0 \leq 1 - \eta_T^+ \leq 99\,11^{-T} \longrightarrow 0, \qquad 0 \leq \eta_T^- \leq 19(9/19)^T \longrightarrow 0.$$

Thus the Bayesian aggregate converges almost surely to 1 under E^+ and to 0 under E^-. The experiments are fixed for all T; their realized states differ.

Both aggregation models also reproduce τ_1 for a second agent using the common garbling

$$G = \begin{pmatrix} 5/8 & 3/8 & 0 \\ 3/8 & 5/8 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Its reports are $(2/5, 3/5, 3/22)$, with equal probabilities on the first two at the respective realized states. Thus both plotted report distributions, the common prior, and the ex ante report probabilities agree across the two aggregation models.