Can AI Agents Agree to Disagree? Aumann's Theorem and the Epistemic Status of Machine Outputs

Revised manuscript. Author information omitted for blind review.

Abstract

Aumann's Agreement Theorem (1976) establishes that two Bayesian agents with common priors and common knowledge of each other's posteriors cannot agree to disagree. The theorem formalizes an assumption that we all commonly rely on: when two people who respect each other's judgment disagree, we assume it must be for a reason. Either one knows something the other doesn't, or they're working under different assumptions, or one of them has made a mistake. And when they agree, we grow more confident that what they agree on is true. Multi-agent systems built on large language models import this assumption into their protocols for debate, evaluation, and consensus. This paper asks what the divergence and convergence of model outputs actually tell an observer. We distinguish three questions: does a model possess beliefs, do its outputs report those beliefs, and can its outputs serve as evidence. We argue that the Aumannian analogy fails at the second question, not the first. Whatever the status of a model's internal states, its outputs cannot be presumed to report them, because nothing in training makes a model's outputs track its internal states the way sincerity norms make a person's assertions track their beliefs. The third question remains open. We answer it by separating the different reasons model outputs can diverge: because the agents hold different beliefs, because they have been exposed to different information, because they were built differently, because of the randomness of sampling, or because they were instructed to argue opposite sides. For each, we ask when an observer should treat the divergence as evidence. When the only difference between two outputs is the randomness of sampling, their divergence says nothing about the world; that much can be proved. In every other case, what divergence means depends on which of these causes can be ruled out. Seen this way, AI agents do not hold debates; they manufacture text on command. When they agree, they are often just repeating shared mistakes rather than confirming facts. And when they disagree, they are not revealing hidden knowledge, leaving us to rely solely on their statistical track record. Our analysis applies to autoregressive language models and the multi-agent systems built from them.

1. Aumann's Theorem and the Epistemic Reading of Disagreement

In 1976, Robert Aumann proved that two rational agents who share a common prior probability distribution and who have common knowledge of each other's posterior beliefs about an event must assign it the same probability (Aumann, 1976). They cannot, in the precise sense the theorem defines, agree to disagree.

The result can be stated compactly as follows. Let (Ω, F, P) be a probability space representing a shared prior over states of the world. Let Π_1 and Π_2 be partitions of Ω representing the private information of agents 1 and 2. Each agent i observes which element of Π_i contains the true state ω and updates by Bayes' rule to form a posterior $q_i = P(A \mid \Pi_i(\omega))$ for an event $A \in F$. Aumann proved that if the posteriors q_1 and q_2 are common knowledge, then $q_1 = q_2$.

Three situations need to be distinguished here, because the theorem says something different about each, and this paper will relate to all three. First, two agents may hold different posteriors because they have observed different things; this is fully compatible with Aumann's model and is, indeed, its starting point. Second, one agent may learn the other's posterior and update on it; here the learned posterior functions as a public signal, and the updating it prompts is ordinary Bayesian conditioning, informative but not yet Aumann's full theorem. Third, the posteriors may become common knowledge, each agent knowing the other's posterior, knowing that the other knows hers, and so on; the theorem concerns this situation alone, and its conclusion is that here disagreement cannot persist. The claim that disagreement is informative belongs to the second situation. The claim that it is impossible belongs to the third.

A further distinction is necessary. As a formal result, the theorem concerns probability spaces, partitions, and a common knowledge relation, and it holds under any interpretation of the measures that has this structure; no doxastic reading of the probabilities, no interpretation of them as anyone's beliefs, is therefore required. But what gives the theorem its epistemological significance is an interpretation under which the posteriors are read as the credences of agents, and their assertions as testimony of what the agents have concluded. Under that interpretation, and only under it, the theorem underwrites a familiar habit of reasoning whereby disagreement is taken as a signal that merits a response. Accordingly, if someone who reasons as carefully as I do, and starts from the same assumptions, has reached a different conclusion, I assume she knows something I don't, or that I have erred, and either way her dissent is regarded as evidence that I should take seriously. Peer review, medical second opinions, expert panels, parliaments, and courtroom trials all harness versions of this logic. But disagreement in these practices can also arise for other reasons. Two reviewers reading the same research paper will often identify different issues. They are looking at the very same thing; what differs is what they know, how they think,

and what they notice. Disagreement of this kind is productive as well: each catches what the other would miss. But it is not the kind the theorem describes: Aumann's result concerns only disagreement that signals a difference in information.

This way of reading disagreement, as a sign that the other party knows something or that I have erred, is now being extended to artificial agents. The AI industry is moving from single-model interactions toward multi-agent architectures in which agents built on large language models debate, negotiate, evaluate one another, and reach consensus, with industry forecasts projecting rapid growth in agent-embedded applications (McKinsey & Company, 2025). Research systems use structured debate among LLM agents to inform portfolio analysis (Zhao et al., 2025) and trading decisions (Xiao et al., 2024), evaluation pipelines use models to judge the outputs of other models (Zheng et al., 2023), and consensus among models or samples is widely treated as a reliability signal, in mathematical reasoning, question answering, and the verification of factual claims (Chen et al., 2025; Li et al., 2024; Naik, 2024). All of this depends on the epistemic reading of agreement and disagreement: convergence is treated as confirmation, divergence as scrutiny or as a warning, concession as the recognition of a stronger argument.

This paper asks what the divergence and convergence of LLM outputs should actually be taken to mean. This appears to be an important and consequential question, since AI agents are already part of how we operate in our day-to-day. They contribute to the extended environment in which we form beliefs, decide, and act, with millions of people now beginning their information gathering, their writing, and their decisions in a chat window. What should they make of a situation in which a model disagrees with them, or in which several models agree? We suggest that this depends on our stance toward AI answers and what they amount to. Are they to be understood as beliefs, as reports of AI reasoning processes, or as outputs that can serve as evidence in our own reasoning without our committing to any view about what lies behind them? We argue for the third understanding. Whether models have beliefs is a genuinely open question, and nothing in this paper settles it: interpretability research finds internal representations in these models that track whether statements are true or false (Li et al., 2023; Marks & Tegmark, 2024), but whether such representations amount to beliefs depends on what one takes a belief to be, a question on which philosophical accounts differ (Dennett, 1987; Schwitzgebel, 2002; Herrmann & Levinstein, 2025). The second understanding raises a sharper issue: does the model, in its outputs, reflect whatever belief-like states it has? In the human case we take this for granted. Human disagreement is informative because we treat a person's words as expressing what she has actually concluded; sincerity is the default, and the practice of assertion enforces it. For models there is no such default: nothing in how these systems are built and trained ties what they say to what is going on inside

them (Section 3), and it is here, not at the question of belief, that the analogy with human disagreement breaks.

Even if a model's outputs reflect nothing of what is inside it, they can still be evidence about the world. A thermometer's reading is evidence about the temperature because the reading depends on the temperature, not on any belief the thermometer has about it, for it has none; in the same way, a model's output can be evidence about the world, because what a model says depends, through its training data and its context, and increasingly through tools and sensors that put it in direct contact with the world, on how the world is. This brings us back to the paper's question. If a single output can be evidence, what should we make of two outputs that conflict? What the conflict indicates depends on what produced it. “Models disagree” covers several different situations, and we distinguish between these (Section 4), then we consider what, in each case, a human observer can conclude about the world from the divergence in outputs between agents (Section 5). Start with the simplest case. Ask a chatbot a question, then ask it again, and you may get two different answers. Nothing changed in between, not the world, not the model, not the question. The two answers differ only because the model composes its reply with an element of chance, the way two rolls of a die differ. So the fact that the answers disagree tells you nothing about the thing you asked about, only that the process has some randomness in it. Every other case is more complicated, because when two models disagree, several things could explain it. They may have been trained on different material. They may have been given different documents or tools. They may have been told to argue different sides. Or it may again just be chance. What their disagreement tells you depends on which of these explanations you can rule out. Aumann's theorem describes the one situation where everything but a single explanation has been ruled out: his conditions leave a difference in information as the only possibility, and that is exactly why disagreement, in his setting, says so much. In today's multi-agent systems the situation is the reverse. All the explanations are viable at once, and nobody checks which one is the right one. Section 6 returns to the practices described above: debate, consensus, and the model as second opinion, and asks what each is actually doing once the question of where divergence comes from is taken seriously, and what happens when the observer drops out altogether. Section 7 reviews the related research, and Section 8 answers objections.

2. Why the Theorem's Conditions Appear to Hold for LLM-Based Agents

The conditions of Aumann's theorem, common priors, consistent Bayesian inference, and the communication of posteriors, can at first sight seem more naturally satisfied by LLM-based

agents than by humans. After all, two copies of the same model are identical in a way no two humans ever are; different models are trained on much of the same text, whereas humans grow up in different worlds; and in a multi-agent system every agent's output is laid out in full for the others to process, where humans only ever guess at each other's minds. We consider these conditions one at a time. Each analogy turns out to be misleading, and each fails at a different point; the three failure points, taken together, map exactly what separates these systems from Aumann's agents, and the separation is important, because each one removes a different inference that we are accustomed to drawing from agreement and disagreement.

2.1 Common priors

The common prior assumption has been the most contested of the theorem's conditions in its application to humans, whose beliefs form through radically different histories (Harsanyi, 1967–68; Morris, 1995; Gul, 1998; Aumann, 1998). LLM-based agents can seem to do better: two instances of one model share identical weights and hence identical output distributions, and even across model families, heavily overlapping training corpora invite the thought of a shared epistemic foundation. But this approach conflates training data with priors. The analog of a prior, if there is one, is encoded in the parameters, and different architectures, initializations, optimization trajectories, and fine-tuning procedures produce different parameters even on identical data. Overlapping data is an input to the process that fixes the prior, not the prior itself, and the common prior condition fails for any pair of distinct models (whether the training differences that break the analogy are themselves informational, records of differential exposure to the world, is a further question, and a more consequential one; we take it up in Section 4.3.) One qualification is needed: claims about the actual overlap of proprietary training corpora are conjectural, since little is publicly known about the composition of frontier-model training data.

2.2 Consistent inference

Bayesian rationality requires consistency at the level of propositions: an agent who affirms that P is true is thereby committed to affirming that not-P is false, and to giving the same answer in whichever of these two forms the question is put. LLMs often fail this test (Kassner & Schütze, 2020; Jang et al., 2022; Jang & Lukasiewicz, 2023). It might seem that determinism excuses them, since each answer is computed by a fixed function of the input, and a fixed function at least never contradicts itself on the same input. But that is not the consistency that is relevant here: any function is deterministic given fixed inputs, and what rationality requires is not that the same input always produce the same output, but that answers to logically related questions hang

together. One might want to push the test further: surely a rational agent should also give the same answer to the same question however it is worded, and LLMs are notoriously sensitive to wording (Elazar et al., 2021; Sclar et al., 2024). We rest nothing on this stronger test, because it is not clearly a test of rationality at all. Two wordings of a question are rarely just two wordings. How a question is put can itself carry information, for instance about the assumed context, the expected answer, the asker's concern, and an agent that responds differently to different wordings may be responding to that information rather than contradicting itself. To turn wording-sensitivity into a charge of irrationality, one would need a way of saying when two prompts carry exactly the same information, and no such account is available. The clean failures are the logical ones: affirming P while denying that not-P is false is a contradiction on any reading of what the questions carry, and failures of that kind are all the argument needs.

2.3 Communication of posteriors

The theorem requires common knowledge of posteriors, something humans can only approximate. LLM agents seem better placed here. In debate or consensus settings, two models arguing for and against a certain stock trade, say, or several models tasked with diagnosing a radiology scan, every response goes into what we might think of as a shared thread (the accumulating exchange that is fed back to every model as part of its input), and every agent processes it before responding again. Because nothing in such a scenario is undisclosed, common knowledge seems to be a given. And yet, mutually shared text, or data, is not the same as mutually shared knowledge. Common knowledge requires that when one of the models, call it A, adds its assessment to the thread, the assessment convey to model B what A has concluded from its own information. It also requires that B's contributions do the same for A; that A know B takes its contributions this way, and B know the same of A; and so on up. But all a shared thread guarantees is that each of the agents encounters the same text. We cannot know what an agent concludes from that text. The condition fails at this first step, the step Section 3 examines, before anything about who knows what about whom even comes up.

Taken together, a picture emerges in which LLM-based agents are treated as if they satisfied Aumann's conditions: overlapping data standing in for common priors, deterministic computation for consistent inference, a shared thread for communicated posteriors. But this is inaccurate and misleading. If models A and B were trained on overlapping data, it does not follow that they share a prior. If each computes its answers deterministically, it does not follow that either answers consistently: determinism fixes each answer on its own, while consistency is a relation between answers, and nothing ties the model's answer about P to its answer about not-P.

And if both process the same shared thread, it does not follow that either knows what the other has concluded, let alone that they know it in common. The picture has nonetheless been close enough in surface form to ground an implicit premise in several research programs, that model agreement and disagreement can be read the way expert agreement and disagreement are read[i]. The next three sections examine what model agreement and disagreement are actually evidence of.

3. Beliefs, Reports, and Evidence: Three Questions

LLM-based agents produce divergent outputs. Different models give different answers to the same prompt (Yang & Wang, 2026), the same model gives different answers across runs (Atil et al., 2025), and models placed in debate settings argue, revise, and sometimes fail to converge (Du, et al., 2023). So what, if anything, do these divergences tell us? Answering this requires answering three questions: First, does the system possess beliefs, in the minimal sense of proposition-level credences that inform its outputs? Second, are its outputs reports of those states, in the way a person's sincere assertion is a report of what they believe? Third, can its outputs provide evidence to an observer?

The questions are logically independent. A negative answer to the first two does not settle the third: a thermometer has no beliefs and reports nothing, yet its reading is evidence about the temperature in the room. And granting that AI agents have beliefs does not settle the second question, because an AI could have internal truth-tracking states while producing outputs that are decoupled from them, just as a person can believe one thing and assert another. What we want to establish here is where the Aumannian analogy fails.

3.1 The belief question

Aumann's theorem is undemanding. It requires probability assignments over a state space, conditioned on information via partitions, with the probability assignments standing in the common knowledge relation. It does not require consciousness, a vested interest in accuracy, a unified worldview, or verbal reports. A few hundred lines of code, a system nobody would call a 'believer' in any rich sense, could satisfy its conditions in full. What's more, even common arguments against attributing beliefs to machines (e.g., Bender & Koller, 2020; Shanahan, 2024) demand more than what Aumann's theorem does. This is of importance for locating where the analogy fails. If machines lacked beliefs even in the theorem's minimal sense, the analogy would collapse at the first question, and nothing more would need saying. But the usual grounds for

denying machine belief show no such thing. Having a vested interest in accuracy (or a 'stake' in being right) is not a condition on possessing a credence. Frozen weights do not rule out belief revision, since a fixed Bayesian program revises with every new observation while its code and prior never change. Answering differently in a new context does not separate generation from genuine updating, since updating is also new output from new information. And penalizing sensitivity to phrasing presupposes a way of saying when two phrasings carry the same evidence, which no one has. So the belief question is not where the analogy can be shown to fail. If it fails, it fails further along.

What remains is a genuinely contested question. On Dennett's intentional stance (1987), a system has beliefs to the extent that attributing them supports prediction and explanation of its behavior, and LLMs frequently clear that bar. On dispositional accounts (Schwitzgebel, 2002), belief is a matter of patterned dispositions, and the patterns exhibited by current models are extensive but unstable in ways that resist clean verdicts. On representational accounts, the interpretability findings that models encode truth-tracking internal structure (Li et al., 2023; Marks & Tegmark, 2024) bear directly on whether there is anything belief-like for outputs to report, and recent work has begun to articulate standards a representation would have to meet to count (Herrmann & Levinstein, 2025). We take no side.

3.2 The report question

When Mia says that it will rain, her statement is a report. It expresses Mia's assessment of whether it will rain. Mia stands behind her statement, and this practice of stating things is governed by norms of sincerity and accountability. These norms are what make it legitimate for the hearers of Mia's statements to treat Mia's utterance as evidence of what she concluded. This reporting-relation is necessary according to the epistemological reading of Aumann's framework.

Consider now what this means for machine agents. In Aumann's framework, the agents communicate posteriors, and for such communication to take place at all, each agent's output must function as a report of its posterior. Applying the framework to LLM-based agents therefore assumes that their outputs are reports. The same assumption underlies the practices with which we began. A model dissents from your judgment, and you reconsider: you are treating its dissent *as if* it were a peer, or a colleague, disagreeing with you. If several models interacting with each other come to the same conclusion, you may end up trusting that conclusion more because they all agreed. You are relating to their consensus as you would to independent experts converging on

a diagnosis or prognosis. In each of these cases, a generated output is treated as if it were a report of the system's own assessment.

It is this assumption that does not hold. Grant, for the sake of argument, that a model has internal representations that track the truth of propositions, such that, for instance, the statement "it will rain tomorrow" is marked somewhere in the system as likely true or likely false. The path from these internal states to the model's output still goes through a series of distinct processes, including next-word prediction, fine-tuning, and reinforcement learning from human feedback (RLHF), and each of these processes, en route to providing an answer to your prompt, is subject to pressures that have nothing to do with fidelity to the machine's internal state. These include fluency, helpfulness, the avoidance of offense, and the approval of human raters. What finally appears on the screen in front of you answers to all of these pressures at once; fidelity to the internal state is, at best, one consideration among several.

And this isn't only a theoretical possibility. Models have been shown to say things that do not match what is going on inside them. In one set of experiments, researchers gave models multiple-choice questions and rigged the setup so that the correct answer to every practice question was "A." The models picked up the pattern and started answering "A" on new questions too, even when "A" was wrong. Then the researchers asked the models to explain their reasoning step by step. The explanations never mentioned the pattern. Instead, the models offered ordinary, plausible-sounding justifications for each answer, as if the choice had been made on the merits, when it was the pattern, not the merits, doing the work (Turpin et al., 2023). In a second set of experiments, models answered questions correctly and the user then pushed back. The models often backed down and changed their answers. They didn’t do this because anything inside them had changed, but because they learned that agreement gets rewarded (Sharma et al., 2023). In both cases, what the model said answered to something other than its own state: to the planted pattern in the first, to the user's approval in the second.

Hence a model stating that "the probability of rain is 70%" may do so because that string of words is a probable continuation in context, and not because some state of the system assigns 0.7 to the proposition being true. And nothing in how these systems are built and trained enforces this correspondence between internal state and external output. The point here is not that models never make statements that are true. They do, often. The point is that the connection between what a model says and what is going on inside it is structurally unreliable and, at present, unverifiable, whereas human statements are held to sincerity as a norm of assertion. If the reporting relation cannot be presumed, the inferences we draw from model

outputs, dissent as peer disagreement, consensus as expert convergence, lose their license, and this is the exact place where the Aumannian analogy breaks. Agents cannot be presumed to communicate what, if anything, they believe.

3.3 The evidence question

Even if model outputs cannot be considered reports of internal states, they can still provide us with evidence about how things are in the world. In the same way a thermometer tracks temperature without *believing* anything about it, a model can track how things stand without reporting its own or anyone else's beliefs. In this sense, all that is needed for the thermometer's reading to count as evidence is that it varies with the state of the world. Think medical tests, seismographs, and statistical classifiers that work like this. And this approach can also hold for language models. A model's training data consists of observations, testimony, and records of the world, while its immediate context focuses that broad empirical base onto the specific question at hand. Where the output depends on these, the causal chain runs from the world through data and context to the output, allowing an observer to draw inferences about the world. Outputs can be evidence without being reports.

Evidence functions in two fundamentally distinct ways. When you rely on Mia's claim that it will rain, you are relying on her epistemic agency: you trust her as a competent evaluator of the data, and the fact that she has reached a conclusion becomes your evidence. When you rely on a thermometer, however, there is no underlying assessment that the thermometer does that you need to trust. What you rely on instead are facts about the thermomenter qua instrument: how well calibrated it is, how often it is wrong, the conditions under which it fails. Evidence derived from testimony is mediated by judgment; evidence derived from an instrument is underwritten by a track record of reliability.

Here one might object that we trust Mia's forecasts for the same reason we trust the thermometer's readings: because of her track record. That is true, but the two records vouch for different things. Mia's record certifies her judgement, and what her track record licenses you to rely on is her judgment: that she, having considered the relevant evidence, concluded that *it will rain*. The thermometer's record certifies no such judge; it merely establishes the physical link between its readings and the world.

The prevailing approach to machine judgment rests on a fundamental category mistake: it treats instrument readings as testimony. When we read a model's output as the thought-out view of a peer, we borrow the epistemic convenience of how we respond to humans: we simply take the

speaker at their word. Treating a model as an instrument, by contrast, demands rigorous epistemic labor. It requires us to calibrate the system, map its error distribution, and measure how its failures correlate with those of other models. This lets us discount its readings in domains where it is brittle, trust it when it operates within its calibrated bounds, and recognize when the agreement of a second model is a genuine confirmation rather than the same error echoed twice, distinguishing, for instance, between independent instruments whose converging readings actually constrain the world, and correlated systems whose shared architecture simply echoes the same underlying blind spot. It is unsurprising that the industry has defaulted to the illusion of testimony, as it demands no such effort. But this work is imperative: an instrument's statistical track record is the only genuine evidence these systems can offer.

We argue that anyone relying on AI today must adopt this instrumental stance. Once we stop treating models like peers and start treating them like tools, it completely changes what it means when we get conflicting answers from different models. If model outputs are merely perceived as instrument readings, a divergence between them ("disagreement") forces us to identify the causal source of the discrepancy. We need to ask whether the instruments ingested different information, whether they have different system architectures, or whether they are perhaps just exhibiting the random noise of stochastic sampling. These are completely different structural phenomena which our inherited doxastic vocabulary would mistakenly label as "disagreement." In practice, when it comes to AIs interacting with each other, this so-called disagreement manifests not as a clash of minds, but as a spectrum of mechanical artifacts: stochastic variance across sampling runs, forced adversarial search in debate protocols, or the uncalibrated cross-measurement of one model evaluating another. The next section replaces this blanket term of 'disagreement' with a precise taxonomy, establishing exactly what kind of evidence each form of divergence actually provides.

4. A Taxonomy of Output Divergence

The preceding section argued that the Aumannian analogy fails at a specific point: LLM outputs are not, in general, reports of posterior beliefs, and the inferences modeled on human peer disagreement therefore cannot be transferred to model outputs automatically. But this conclusion settles less than it might appear to. It establishes only that a particular family of inferences lacks its usual epistemic license; it does not entail that divergence among outputs carries no information at all. Whether it does depends entirely on *why* the outputs diverged, an inquiry that fractures into several distinct causal scenarios. It is tempting to sweep every machine case under a single label

like "pseudo-disagreement," but doing so conflates phenomena with entirely different causal structures and evidential profiles. And so, before asking what divergence tells an observer, we must first categorize the different forms it can take. We distinguish five.

4.1 Doxastic disagreement

To establish a baseline, consider a scenario where two agents hold different credences in the same proposition. This represents "doxastic" disagreement, a genuine clash of internal beliefs, which is the core case addressed by Aumann's theorem and the peer disagreement literature. In this scenario, each agent has weighed the evidence and arrived at a contradictory conclusion. When two physicians examine a patient and reach different diagnoses, each diagnosis expresses a conclusion the physician has drawn and stands behind. Disagreement of this kind is informative because conflicting conclusions indicate that something upstream, such as evidence or reasoning, differed. Whether machine systems ever instantiate doxastic disagreement depends on the question of machine belief discussed in Section 3, and nothing in what follows requires settling this issue. We include this category simply as a theoretical anchor, a standard of true disagreement used to evaluate the machine behaviors that follow.

4.2 Informational disagreement

In cases of informational disagreement, divergence arises not from conflicting beliefs, but because two agents have been exposed to different information. A radiologist who has seen a patient's prior scans may read the current one differently from a colleague who hasn't, and the difference between their readings directly reflects this informational asymmetry. Crucially, informational disagreement can occur whether or not the doxastic question is settled. Because it is defined entirely by the causal source of the divergence, differential exposure to information, rather than the mental status of the conclusions, this category applies directly to contemporary multi-agent systems.

Applying this framework to contemporary LLM-based agents requires drawing the proper unit of analysis. The relevant entity is not the underlying language model itself, but the complete agent architecture: the model integrated with its prompt, context, memory, retrieved material, tools, and interaction history. At this level, two agents built on the exact same base model can occupy entirely different informational vantage points. The AlphaAgents system (Zhao et al., 2025), an experimental multi-agent framework developed by researchers at the global asset manager BlackRock, illustrates this perfectly. Its three agents share a base model but are informationally partitioned: the valuation agent receives price and volume data alongside

computational tools, the sentiment agent receives financial news, and the fundamental agent receives corporate disclosures via retrieval pipelines. When these agents produce divergent assessments of an equity, the discrepancy reflects the fact that each has been exposed to material the others have not seen. This is differential information in a strictly Aumannian sense, regardless of what we conclude about machine belief. To locate the source of their divergence entirely in their system prompts is to fundamentally misdiagnose the causal structure of the disagreement.

4.3 Procedural disagreement

Whereas informational disagreement stems from the data a system receives, procedural divergence arises from how the system is built. Two models may produce entirely different outputs because they rely on distinct architectures, initializations, optimization trajectories, fine-tuning procedures, prompts, or decoding methods. In these instances, the divergence traces exclusively to differences in *how* the systems process their inputs, rather than differences in *what* they are given.

The difficult case is training data. It is natural to classify all differences in training as generative divergence, a difference in the function mapping inputs to outputs rather than a difference in information. But training corpora are not arbitrary function-fixing material. They consist of observations, testimony, measurements, and records of the world, and two models trained on different corpora have, in effect, seen different things. A model whose training data includes the relevant scientific literature stands in a fundamentally different informational relation to a question than a model whose data does not. Think of a specialized medical model trained on a decade of randomized controlled trials versus a general-purpose model trained on internet forums: the former has not merely internalized a different statistical mapping, it has effectively observed clinical realities the latter has never encountered. While this differential exposure to the world is ultimately baked into the model's parameters rather than stored in anything resembling human memory, its causal origin is informational, and it must be evaluated as such.

The boundary between informational and procedural divergence, therefore, does not map neatly onto the boundary between humans and machines, nor between prompt context and underlying weights. Cross-model divergence must be classified case by case: it is informational to the extent that the architectural differences encode differential exposure to the world, and procedural to the extent that it reflects differences in how comparable exposure was processed. In practice, these two causal strands are deeply entangled. For proprietary models, this entanglement is usually impenetrable from the outside. This epistemic opacity, the observer's inability to reliably attribute

a divergence to either unique information or unique processing, sets the stage for the structural vulnerabilities explored in Section 5.

4.4 Sampling variation

When informational inputs and procedural architectures are held perfectly constant, divergence can still arise from the sheer randomness of stochastic sampling. In these instances, nothing about the world differs between runs; nothing about the model differs; nothing about the input differs. The divergence is generated entirely by the mechanical randomness of sampling from a fixed conditional distribution. This makes it the conceptually cleanest case in the taxonomy, and the single instance where divergence genuinely conveys zero information about the world, a limitation that Section 5 establishes in formal scope.

4.5 Divergence by design

Unlike divergence that emerges organically from data or architecture asymmetries, divergence by instruction occurs when systems are explicitly commanded to defend conflicting positions. Formal AI debate protocols are the central example here: two models are pitted against one another to expose the flaws in a given claim. Each agent generates text arguing for its assigned side, and the resulting conflict is a designed feature of the protocol, not a discovery about either system's internal assessment of the issue at hand. While this assigned opposition is not disagreement in any doxastic sense, it is not epistemically inert. Even though the agents are only role playing, the debate they produce can still yield genuine evidence. The human parallels of institutionalized dissent demonstrate its epistemic value: legal advocacy, structured debate, red-teaming, and the assignment of a devil's advocate all use artificial opposition to surface arguments, objections, and weaknesses that unopposed inquiry tends to miss. What these practices show is that assigned opposition can be highly truth-conducive as a mechanism of search and error elicitation. What they do *not* license is treating the opposition itself as evidence of conflicting assessments. In this scenario, the two functions of debate, acting as a tool for discovery versus signaling a genuine conflict in information, are entirely decoupled. Keeping them separated is precisely what the inherited doxastic vocabulary of “agreement” and “disagreement” makes so difficult.

4.6 What the taxonomy replaces

When we encounter opposing arguments, our conversational habits naturally assume a thinking mind behind them. This reflex tempts us to view disagreement as a strict binary—either a genuine

clash of internal beliefs on one side or a superficial, empty conflict on the other. This taxonomy replaces the rigidity of that binary approach with a spectrum, ordered roughly by how much of the divergence traces back to differences in information about the world. Doxastic and informational disagreement sit at one end, and stochastic divergence sits at the other. Procedural divergence ranges over the middle depending on what the constructional differences encode. Divergence by instruction sits off the spectrum entirely, because it is not a form of disagreement at all, but rather a form of search guised in disagreement's clothing.

For practical purposes, the important question is not where a given divergence sits on this spectrum in itself, but what an observer confronted with that divergence is entitled to conclude. That question is observer-relative and it is the subject of the next section.

5. When Is Divergence Evidence?

5.1 Setup

We model the situation of an observer confronted with the outputs of two LLM-based agents. We formalize this setup through the following core components.

Let W represent the objective state of the world (which remains external to the agents), and let x be the input given to the agents. Each agent $i \in \{1,2\}$ is a pair (f_i, I_i), where f_i is a generative function and I_i is the agent's information channel: its context, retrieved documents, tool outputs, and memory. Agent i's output is a random variable Y_i distributed according to $f_i(\cdot \mid x, I_i)$, a conditional distribution over strings.

Yet strings are not the right level at which to define disagreement. Two strings can differ while expressing the same answer, and identical strings can be produced by very different underlying distributions. We therefore introduce a semantic mapping σ from output strings to a space A of answer classes, and define all notions of agreement and disagreement at the level of $\sigma(Y_i)$ rather than Y_i. How σ is specified will vary with the task: for a classification question, A is the label set; for a numerical estimate, a partition of the reals; for open-ended tasks, σ must be supplied by an evaluation procedure, and disagreement is only well defined relative to it. This relativity is not a defect of the model. It reflects the fact that whether two outputs 'disagree' depends entirely on what the outputs mean, not how they are phrased.

Finally, let O be an observer with a prior over W and with background knowledge K about the agents: which models they are, what channels they have, and whatever else O knows about the

setup. The observer may be a human user, a judge model, or a downstream system; nothing in the analysis depends on which.

We can now separate two notions that are easily conflated.

Definition 1 (Genuine disagreement). As in Section 1: agents sharing a common prior P over Ω, with information partitions Π_1 and Π_2, genuinely disagree about event A at state ω if $P(A \mid \Pi_1(\omega)) \neq P(A \mid \Pi_2(\omega))$. In the taxonomy's terms, genuine disagreement is doxastic disagreement with an informational source.

That is, for a conflict to be considered genuine, two conditions must be met: the agents must hold conflicting internal assessments of the truth (precluding the *divergence by instruction* discussed earlier), and this conflict must stem from a true asymmetry in their underlying information about the world.

Definition 2 (Distributional divergence). Agents 1 and 2 diverge distributionally on input x if the pushforward distributions of their outputs through σ differ: $\sigma \circ f_1(\cdot \mid x, I_1) \neq \sigma \circ f_2(\cdot \mid x, I_2)$. This is a fact about the two generative processes, and it can hold whether or not any outputs have been sampled.

In other words, the divergence is an inherent property of the agents' statistical wiring, rather than a mere fluke of a specific generation. This misalignment exists as a latent state within the models, independent of whether an observer ever prompts them to manifest those answers as text.

Definition 3 (Sample divergence). Realized outputs s_1 and s_2 exhibit sample divergence if $\sigma(s_1) \neq \sigma(s_2)$. Sample divergence can occur without distributional divergence, as when two draws from the same distribution land in different answer classes, and distributional divergence can hold while particular samples happen to agree. Neither notion reduces to the other, and the arguments below keep them separate.

In short, while distributional divergence is a fact about the models' internal wiring, sample divergence is merely a fact about what they happened to generate *this time*. It is the surface-level clash of realized text.

5.2 The observer criterion

Let D denote the event that $\sigma(Y_1) \neq \sigma(Y_2)$: the observer sees the two agents disagree, in the semantic sense. The question the rest of this section will pursue is when D can serve as evidence.

Definition 4 (Evidential relevance). The event D is evidentially relevant to W for observer O if and only if

$$P(W \mid D, x, K) \neq P(W \mid x, K).$$

Divergence is evidence about the world exactly when seeing it should change what the observer believes about the world, given everything else the observer knows. Three features of this criterion deserve emphasis. It is observer-relative: the same divergence event can be informative for one observer and uninformative for another, because K is part of the condition. It is symmetric in agreement and disagreement: the logical inverse of D, observed agreement, is evidentially relevant under the same criterion, a fact we return to in Section 6. And it detaches the evidential question from the doxastic one: nothing in Definition 4 mentions belief. Whether the agents believe anything is simply not what the criterion tests, which is why a negative answer to the belief question, however well established, cannot by itself establish that divergence is uninformative.

5.3 Sample divergence is uninformative for a knowledgeable observer.

We now state the sense in which sampling variation is genuinely uninformative.

Proposition 1 (Conditional independence of sampling variation). Let Y_1 and Y_2 be independent draws from a single fixed distribution $f(\cdot \mid x)$, with identical information channels folded into x and no external information sources. Let the observer's knowledge K include the identity of the model and the input, so that O knows the predictive distribution $p = \sigma \circ f(\cdot \mid x)$ over answer classes. Then (Y_1, Y_2) is conditionally independent of W given (f, x), and consequently

$$P(W \mid D, f, x) = P(W \mid f, x).$$

Argument. The world influences the outputs through exactly two channels: the process that produced f (training) and the input x. Conditional on f and x, both channels are fixed, and the joint distribution of (Y_1, Y_2) is $p \otimes p$ regardless of the value of W. In particular the probability of divergence, $P(D \mid f, x) = 1 - \Sigma_a\, p(a)^2$, is a function of p alone. An event whose probability does not vary with W cannot shift a rational observer's credence in W.

This proposition formalizes the claim that sample divergence (the fourth category of the taxonomy) tells an observer nothing about the world. Everything the world contributed to the two samples, it contributed through the training and the prompt. Because those inputs are identical across the samples and already conditioned on, the divergence itself carries no new worldly information. However, this formalization is subject to two strict boundary conditions that require immediate clarification.

First, the result is relative to an observer who knows the predictive distribution. An observer who knows which model and which prompt but not how the model's probability mass is spread across answers does learn something from D, namely something about the entropy of p. That is information about the model, not about the world, but it is information, and Proposition 1 does not deny this.

Second, entropy is not sealed off from the world when aggregated across multiple tasks. Across many inputs, the spread of a model's predictive distribution correlates with properties of the inputs: their difficulty, their ambiguity, the quality of the evidence bearing on them, and the model's probability of error. An observer using sample divergence to flag inputs for review is therefore reasoning soundly; this familiar diagnostic use of divergence is derivable from within the model. But the inference has a specific structure, running from divergence to predictive entropy to expected error, not from divergence to unshared evidence. What the observer learns is that the model is unreliable here, which is grounds for reducing confidence in the answer. What the observer does not learn is that one of the runs knows something the other doesn't, because there is nothing of that kind to learn.

5.4 The general case: attribution

Outside the conditions of Proposition 1, the conditional independence generically fails, and divergence *can* be evidence. A simple causal structure makes this clear:

$$W \rightarrow \text{data and observations} \rightarrow \text{parameters and context} \rightarrow \text{outputs}.$$

Whenever the two agents' positions in this chain differ in a way that depends on W, or whenever divergence is statistically associated with error, difficulty, or evidence quality given the observer's knowledge, D satisfies Definition 4. Three configurations cover most systems. When agents have different retrieval documents, tools, or context, as in the AlphaAgents architecture, their divergence can reflect differential information in nearly Aumann's sense, and it is evidence of roughly the kind human expert disagreement provides, subject to the reliability of the agents. When separately trained models diverge, the divergence is evidence to the degree that the training differences encode differential exposure to information bearing on the question, which is the case-by-case classification Section 4.3 described. And when the agents differ only in decoding method or other processing choices, with comparable informational exposure, divergence approaches the condition of Proposition 1 and carries correspondingly little information about the world.

The observer's problem, in every case, is attribution. Divergence is a symptom with multiple possible causes, informational difference, constructional difference, sampling noise, assigned roles,

and its evidential force for a given observer depends on which causes the observer's knowledge K can rule out. This reframing gives Aumann's theorem its proper place in our argument. Aumann's setting is the limiting case in which attribution is complete: common priors eliminate constructional difference, Bayesian rationality eliminates processing noise, and common knowledge of posteriors eliminates sampling. The only cause left standing is differential information, which is why, under the theorem's conditions, disagreement is maximally informative and cannot even persist. Multi-agent LLM practice generally sits near the opposite end. The candidate causes are confounded, the observer's knowledge rarely reflects the differences between them, and the evidential force of raw agreement or disagreement is accordingly weak. Weak, however, is not zero, and the difference between the two is important when it comes to system design: it is the difference between discounting the import of a signal and discarding it completely.

5.5 Two senses of "epistemic uncertainty"

It is tempting to map genuine disagreement onto epistemic uncertainty and machine divergence onto aleatoric uncertainty. But this would be wrong. Understanding why it is wrong will clarify the argument.

In machine learning, disagreement within an ensemble of models is a canonical proxy for epistemic uncertainty, that is, for uncertainty about the model: it can indicate underspecification, distribution shift, or lack of training support, and it is used on exactly this basis in methods such as BALD (Houlsby et al., 2011), ariation across separately trained models, or across training runs, is model uncertainty in this standard sense (Lakshminarayanan et al., 2017), and calling it aleatoric misdescribes it. The aleatoric characterization fits only the fourth category of the taxonomy, variation under stochastic decoding from a fixed model. Even there it must be handled carefully, since, as Section 5.3 showed, such variation reveals the spread of the predictive distribution, which an observer can put to use in modern semantic uncertainty frameworks (Kuhn et al., 2023).

What our argument needs is not the aleatoric/epistemic distinction but a distinction between two senses of "epistemic" that the two literatures attach to the same word. In machine learning, epistemic uncertainty is uncertainty about the model, the kind that more evaluation of the model would reduce. In the Aumannian setting, the epistemic content of disagreement is unshared evidence about the world, held by rational agents. Ensemble disagreement is epistemic in the first sense without thereby being epistemic in the second. A system designer who observes divergence among model instances and concludes that the models hold conflicting evidence has moved from the first sense to the second without an argument, and this slide, rather than any confusion between chance and ignorance, is the interpretive error our framework identifies.

6. Implications: Reading the Practices Through the Framework

The preceding sections replaced a binary classification with a nuanced taxonomy of divergence, a criterion for assessing evidential value, an attribution framework to measure that value, and a sharp distinction between treating machine outputs as testimony versus instrument readings. This section applies these tools to the real-world practices that motivated this paper, such as the growing reliance on LLMs-as-judges and the industry's habit of cross-checking models against one another under the assumption that their agreement guarantees accuracy. In each case, a consistent pattern emerges: the practice holds real value, but this value is driven by a mechanism entirely distinct from how it is conventionally described. This misalignment between perceived function and actual mechanics has direct consequences for multi-agent architecture and epistemic safety infrastructure.

6.1 Multi-agent debate: search, not deliberation

The multi-agent debate framework has been proposed as a mechanism for improving reasoning and alignment (Irving et al., 2018; Du et al., 2024; Khan et al., 2024). Within this framework, agents take opposing positions, and a judge (human or AI model) is tasked with identifying the stronger case. In the terminology of our taxonomy, this debate creates opposition by design. It is therefore tempting to conclude that because the agents' conflict is structurally assigned rather than rooted in genuine conviction, the resulting exchange lacks the epistemic substance its adversarial framing ('debate') suggests. We believe this conclusion is too strong, however, and that the courtroom analogy often invoked in this context (e.g., Bandi et al., 2024), likening AI agents to courtroom lawyers, demonstrates precisely the opposite. In court, opposing counsel do not have to believe their own briefs, nor are they required to; and adversarial dynamics are truth-conducive because engineered opposition is an effective mechanism for surfacing arguments, objections, and weaknesses that the uncontested generation of monolithic architectures misses. The mechanism does not rely on the participants' beliefs, but on the exhaustive exploration of the argument space and the competence of the tribunal.

Read this way, debate between models is a search procedure with a verification step, and its epistemic value is instrumental rather than intrinsic. This instrumental value rests on three conditions: the protocol must actually widen coverage, the judge must be capable of distinguishing stronger from weaker arguments, and checking an argument must be easier than constructing one—the wager at the heart of the original debate proposal (Irving et al., 2018). None of these conditions involves the debaters believing anything, and none is undermined by our analysis. Nor

should the stochastic diversity that drives this search be ranked as a weaker epistemic resource than deliberation among believers. The comparison is task-relative: where candidate solutions can be generated cheaply and verified reliably, many samples with a checker can outperform a small group of deliberating experts, and nothing in our framework says otherwise.

The danger our framework highlights lies instead in how these mechanisms are *interpreted*. The vocabulary of debate smuggles in inferences that belong strictly to deliberation: that a concession shows the model yielded to superior logic, that agreement after several rounds means a claim survived rigorous testing, and that failing to push back means no rebuttal exists. On the search reading, a concession is a shift in the conditional distribution given a new context, convergence is merely an artifact of the generation dynamics, and an unrebutted argument may simply be one the sampling never reached. Recent empirical work confirms this exact vulnerability, showing that when models converge on a stance during debate, this alignment stems as much from mere conformity as from genuine updates (Hao et al., 2026), revealing the preceding exchange to be an illusion of conflict rather than a genuine epistemic disagreement. Furthermore, models demonstrate conformity pressure that is entirely independent of argument quality (Zhu et al., 2025). Treating the dynamics of the protocol as if they carried deliberative meaning is the error, and it is an error about attribution, not about the usefulness of the protocol.

6.2 Consensus and LLM-as-judge: correlated instruments

When three human experts independently agree on a diagnosis, their agreement becomes strong evidence, the strength of which arises from their independence. Each experts's assessment is a separate, independent journey from evidence to conclusion. Consensus methods for language models borrow this logic (Zheng et al., 2023; Chen et al., 2025; Li et al., 2024; Naik, 2024): agreement among models, or among samples, is treated as confirmation, and disagreement as a warning. For instance, in an automated triage pipeline, if three distinct models are fed the same patient history and all output a diagnosis of acute appendicitis, the system assigns a high probability to that outcome. If they diverge (e.g., one suggesting appendicitis, another an infection, and a third a kidney stone) this variance acts as an automatic flag, pausing the workflow for human intervention.

One tempting conclusion here is that model agreement is an artifact of shared training data and architectural design, and carries no evidential weight. Under this view, because modern models rely on nearly identical foundational corpora and convergent fine-tuning paradigms, their consensus is indistinguishable from a single homogenized worldview merely echoing itself. The instrument model of Section 3.3, however, supports a more precise articulation of this view.

Agreement among instruments is indeed evidence, but its strength is governed by the conditional independence of their respective failure modes. Three identically miscalibrated thermometers from the same factory line will perfectly agree on a false temperature; conversely, when three fundamentally distinct measurement tools converge on the same reading, that consensus provides robust, independent evidence of the truth.

Models trained on overlapping corpora with similar architectures and similar fine-tuning objectives lean heavily toward the paradigm of the miscalibrated thermometers: their errors are deeply correlated through their shared informational and architectural heritage. Consequently, their agreement is worth far less than the agreement of three independent human assessors, but it still carries more epistemic weight than a single isolated sample, as slight variations in initialization, stochastic decoding, or data shuffling still filter out purely idiosyncratic noise.

Cross-model agreement is therefore *discounted* evidence, with the discount rate strictly set by the degree of dependence. In simpler terms, the more the models share the same underlying structure, the less their agreement is actually worth. The practical failure of current consensus methods is not that they use agreement as a signal, but that they use it undiscounted, treating highly correlated outputs as independent verification, without an estimate of the error correlation that determines what the signal is actually worth. This is akin to building an investment portfolio out of five different tech equities and believing you are fully diversified, ignoring that they are heavily correlated and vulnerable to the exact same systemic shocks.

The danger of this structural blind spot is most obvious in standard reasoning tasks: if three different language models all fail a logic puzzle by offering the exact same answer, an undiscounted consensus method would classify that correlated error as verified truth. However, if an observer knows all three models share the same flawed heuristic embedded in their pre-training data, the evidential discount rate approaches 100%. Their unanimity is not a robust multi-agent consensus; it is simply the same structural bias firing three times.

However, treating all cross-model agreement as mere structural bias oversimplifies the epistemic reality in two crucial ways. First, there is now substantial evidence that separately trained models converge on similar internal representations, plausibly because they approximate shared latent structure in the data-generating process (Huh et al., 2024). This complicates any blanket claim that cross-model agreement reflects only shared construction because to the extent that convergence tracks structure in the world's data, agreement across models and training runs can be evidence about the world, in just the way Section 5.4's causal chain allows. That being said, it remains true that such agreement lacks the evidential profile of agreement among assessors with independent

evidence, and estimating which contribution dominates in a given case is an open empirical problem, but that very possibility changes the default response from dismissal to discounting. Second, the LLM-as-judge paradigm should be read in the same way: the judge is not an authority rendering a verdict but an instrument measuring other instruments, along with its own error profile, a profile whose errors are typically correlated with those of the systems it evaluates, since judge and judged are often built from the same structural foundations, data, and objectives. A judge whose errors are independent of the systems it scores adds real information; a judge that shares their blind spots ratifies them. The design question, again, is the correlation structure, a factor that is routinely overlooked.

6.3 The second opinion: peers and instruments

The epistemology of peer disagreement deals with the problem of genuine epistemic friction: what a rational agent ought to do when an equally competent and well-informed person arrives at a contrary conclusion (Christensen, 2007; Elga, 2007; Kelly, 2010; Konigsberg, 2013). The everyday practice of treating a model as a sounding board, and taking its answers as reasons to update our own, imports this framework. One ground to reject doing so is the broad claim that models have no beliefs: if a system lacks the capacity to hold an epistemic state, it cannot offer the kind of genuine rational evaluation necessary to challenge our own. However, a much firmer objection lies in what is actually required for dissent to carry weight. What makes a peer's dissent evidence is that it reports their *assessment*. A competent evaluator of the shared material has reached a different conclusion, and this fact in itself serves as second-order evidence regarding the correct evaluation of the first-order evidence. A model's dissent cannot be presumed to report an assessment, whatever is true of its internal states, because the reporting relation between internal state and output is exactly what Section 3.2 showed to be unreliable. What a model's dissent gives you is an instrument reading, and the rational response to a discrepant instrument is neither conciliation nor steadfastness but calibration: how reliable is this instrument on questions of this kind, and how correlated are its failures with yours? On some questions the honest answer makes the reading weighty, on others worthless, and the peer framework appears to be the wrong tool for telling which is which.

Knowing the model is an instrument rather than a peer doesn't remove the threat to our independent judgment. Rather, this realization merely exposes our vulnerability to the model's presentation. When a model disagrees with us, it perfectly mimics a confident human expert. By generating flawless prose, structured arguments, and an unwavering tone of certainty, it hijacks our social instinct to defer to perceived authority (Costello et al., 2024; Bai et al., 2025;

Hackenburg et al., 2025). Yet underneath its persuasive delivery, the model remains a statistical instrument, meaning its authoritative style is entirely disconnected from the actual evidental strength of its underlying data. This creates a dangerous mismatch where users systematically abandon their own convictions based on the system's tone rather than its substance. Ultimately, model outputs are packaged in a form that demands a level of trust their automated production simply cannot underwrite.

6.4 Agent-to-agent systems: attribution without observers

The practices above keep a human in the position of observer. A further set of systems removes the human from the point of interaction entirely: agents negotiating with agents, adjudicating between agents, and reaching consensus with agents. While currently concentrated in domains like trading and portfolio research (Xiao et al., 2024; Zhao et al., 2025), these multi-agent architectures are poised to expand into higher-stakes sectors, such as defense, national security, medical diagnostics, and the management of critical digital infrastructure. These remain, at present, research prototypes and experimental systems, but the concern here is with the trajectory and the design approach.

The AlphaAgents system (Zhao et al., 2025) is again the instructive case. The system's three agents occupy genuinely different informational positions, as Section 4.2 described, so its debate protocol is not only re-sampling; divergence between the valuation agent and the fundamental agent can reflect a real asymmetry in what each has processed, and a protocol that surfaces and reconciles such asymmetries is harnessing informational disagreement, the second category of the taxonomy, not the fourth. The residual criticism is that the system treats convergence as the termination condition and the converged answer as thereby validated, an inference that fails to apply the necessary evidential discount. The three agents share a base model, and with it whatever blind spots, biases, and failure modes the base model contributes, so their eventual agreement is the agreement of correlated instruments, which is worth something, but less than the protocol's design assumes, and less than the same convergence among independent systems would be worth.

As Section 5.4 argued, every inference drawn from agreement or disagreement is an exercise in attribution, the diagnostic work of determining whether that alignment stems from independent evidence or shared structural bias, which must be performed by an observer with knowledge of the systems, their channels, and their dependencies. In human-supervised settings that work is at least possible, even if rarely done. Yet in closed agent-to-agent loops, no observer is positioned to perform it. The downstream agent takes convergence as validation and divergence as noise or dispute, leaving the system architecturally blind to the question of attribution. Ultimately, this

vulnerability stems from the complete absence of an attributor. And this risk grows as the systems gain autonomy and reach, because this problematic approach of treating convergence as confirmation is being reproduced at every scale

7. Related Work

The application of formal epistemology to AI systems remains underdeveloped relative to its importance. The relevant work falls into several bodies of literature, and part of this paper's aim is to connect them.

On the formal side, Aumann's original result (1976) and its extensions (Geanakoplos & Polemarchakis, 1982; Milgrom & Stokey, 1982) established the foundations, with later work extending the impossibility of agreeing to disagree to approximate agreement and bounded settings (Monderer & Samet, 1989; Aaronson, 2005). To our knowledge, no prior work applies this framework to LLM-based agents.

On the multi-agent AI side, Irving et al. (2018) proposed debate as a safety mechanism, Du et al. (2024) showed that multi-agent debate can improve factuality and reasoning, and Khan et al. (2024) studied debate as scalable verification—a method for humans to authenticate AI outputs on tasks too complex for direct inspection. Alongside this, recent work has begun cataloguing the failure modes of multi-agent systems (Cemri et al., 2025; Du, et al., 2023) with the aim of exposing how interacting models can reinforce shared blind spots rather than independently converging on factual accuracy. These papers evaluate protocols empirically. Our contribution here concerns the epistemic *interpretation* of what the protocols do, and, as Section 6.1 argues, our framework is compatible with their positive findings while dissenting from the deliberative vocabulary in which the findings are often framed.

On machine belief, our argument intersects with a philosophical literature that formal treatments of agreement have largely overlooked. Whether systems without consciousness or stakes can hold beliefs is addressed by interpretationist accounts (Dennett, 1987), dispositional accounts (Schwitzgebel, 2002), and, most directly, recent work articulating standards for belief representations in LLMs (Herrmann & Levinstein, 2025), which connects to interpretability findings that models encode truth-tracking internal structures (Li et al., 2023; Marks & Tegmark, 2024). Our position, developed in Section 3, is deliberately agnostic on the belief question and locates the failure of the Aumannian analogy in the reporting relation instead, for which the evidence on unfaithful chain-of-thought (Turpin et al., 2023) and sycophantic output shading (Sharma et al., 2023) is central.

On testimony and evidence, work on AI testimony (Freiman, 2024) and machine epistemology (Floridi & Chiriatti, 2020) asks whether AI outputs can function as testimony. Our instrument model of Section 3.3 offers a different approach according to which model outputs are evidence in the way instrument readings are, governed by calibration and error structure rather than by the norms of assertion. This connects the epistemological question to the machine learning literatures on calibration and predictive uncertainty (Guo et al., 2017), on ensemble disagreement as a proxy for model uncertainty (Houlsby et al., 2011; Lakshminarayanan et al., 2017), and on representational convergence across models (Huh et al., 2024), literatures that supply exactly the quantities, reliability, calibration, and error correlation, that the instrument model says an observer needs.

On peer disagreement, the philosophical literature (Christensen, 2007; Elga, 2007; Kelly, 2010; Lackey, 2010; Konigsberg, 2013) analyzes the rational response to a dissenting peer. Section 6.3 argues that the model-as-second-opinion imports this framework where it does not apply, and that the applicable framework is in fact calibration. The point here being that it is not that the peer literature is wrong but that its subject matter, dissent that reports an assessment of the evidence, is not what a model's dissent can be presumed to be.

8. Objections, and Conclusion

One might object that without definitively settling the belief question, this paper lacks a central thesis. Yet, the thesis does not rest on the belief question. The claim defended here is that the inferences routinely drawn from model agreement and disagreement (e.g., specifically, interpreting dissent as peer disagreement, viewing consensus as confirmation, and treating concession as revision) are licensed by a reporting relation that cannot be presumed to hold. Once this assumption is set aside, what remains is a determinate but entirely different evidential reality: the instrument structure—a framework that current practice almost never engages. The claim is conditional but exact, and its practical upshot, that current designs draw unearned inferences, is intact.

One might object that the instrument model proves too much: since human testimony is also the output of a causal process, a complex sequence of neurobiological and environmental inputs producing verbal outputs, why should humans be treated as epistemic peers while models are reduced to mere instruments? If being a causal mechanism disqualifies a system from peerhood, the critic argues, then human assertion must also be demoted. The difference is not a metaphysical one but normative. Human assertion is embedded in norms of sincerity and accountability: a

sincere asserter's utterance is, by design of practice, so to speak, a report of her assessment. Because deviations carry social and epistemic penalties, hearers are thereby licensed to treat assertions as reports as a default mode of hearing them. No comparable norm governs the pipeline from a model's internal states to its outputs, and the documented decouplings (Turpin et al., 2023; Sharma et al., 2023) show that extending this default to models is epistemically hazardous. This objection does contain a lesson though. Humans can be treated as instruments too, as the literature comparing clinical and actuarial judgment has long done, and there is nothing wrong with asking calibration questions of human experts (such as requiring forecasters to assign exact numerical probabilities to future events or asking physicians to state their subjective confidence percentages). The asymmetry is that with humans the testimonial default is earned, while with models it is borrowed outright and without any collateral.

Finally, one might object that the common prior assumption fails for humans too, so the theorem's conditions are met by no one, human or machine. This is correct, and strictly speaking, the paper treats the theorem as the limiting case which shows what disagreement looks like when it can mean only that the agents know different things (or have different information). No human institution meets the conditions either. Jurors see the same evidence and still divide in the conclusions they arrive at, because they weigh the evidence differently. But in cases like this, one condition holds that the theorem never even bothers to state, because for humans it goes without saying: when a juror announces her verdict, the verdict is hers. So when humans disagree, the disagreement is at least a disagreement between conclusions that people actually hold. That is why more discussion, more disclosure, and more deliberation help: they bring the conclusions, and the reasons behind them, further into the open. With machines, none of this can be assumed. Their outputs may not be anyone's conclusions at all (Sections 3 through 5), and putting models into further conversation with each other brings nothing further into the open. What helps instead is calibration: learning about the systems themselves, how reliable they are, when they fail, how their errors correlate. The objection is therefore right that the theorem's conditions are met by no one, but wrong about what follows. The theorem's value is that it shows where each kind of agent falls short of it. Humans fall short in how they weigh evidence and how they communicate; machines fall short in whether their outputs can be considered to be conclusions at all. These two shortfalls call for different remedies.

Conclusion. Aumann's theorem formalizes the most natural reading of disagreement: when two parties who respect each other's judgment diverge, there is a reason, and when they converge, the convergence is a sign they're right. It is tempting to argue that machine divergence supports

no such reading, because there are no beliefs behind the outputs. The argument of this paper is at once more modest and, we believe, more consequential. Whether there are beliefs behind the outputs is an open issue, and it does not matter for the practices examined here. What matters is that the outputs cannot be presumed to *report* what lies behind them, so the readings that assume they do (peer dissent, pooled assessment, rational concession, etc) are unlicensed. What such outputs can be regarded as is evidence in the way a thermometer's reading is evidence. And they warrant inference in proportion to what the observer knows about their reliability, their calibration, and the correlation of errors between them. For instance, divergence between two samples of one model, on one prompt, tells nothing about the world. Divergence between agents with different information channels can reveal a great deal. And between these poles, everything turns on what led to divergence. The systems now being built and used, systems that debate, judge, converge, and transact with one another, are built with that question largely unasked and model agreement is taken at face value. These outputs are not necessarily empty of epistemic content, for they carry information whose value depends on reliability and error correlation, characteristics which their current designs don't measure. What this paper has tried to do is to relate to large language models as instruments, and to reserve the epistemic deference we show to peers for agents whose statements can be considered reports, presumed to express their conclusions.

Acknowledgements

I would like to thank two anonymous reviewers whose comments substantially influenced this paper. The taxonomy of Section 4 adapts a classification proposed by one of the reviewers. Additional thanks to Robert Aumann for his thoughtful comments.

References

Aaronson, S. (2004). The complexity of agreement. *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, 634–643.

Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., Rajagopal, G. R., Sloan, A., Tudrej, T., Ture, F., Wu, Z., Xu, L., & Baldwin, B. (2025). Non-determinism of "deterministic" LLM system settings in hosted environments. In *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems* (pp. 135–148). Association for Computational Linguistics. arXiv:2408.04667.

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics*, 4(6), 1236–1239.

Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C., & Willer, R. (2025). LLM-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1), 6037.

Bandi, C., & Harrasse, A. (2026). Debate, Deliberate, Decide (D3): A Cost-Aware Adversarial Framework for Reliable and Interpretable LLM Evaluation. *arXiv*. https://doi.org/10.48550/arxiv.2410.04663

Chen, Z., Li, J., Chen, P., Li, Z., Sun, K., Luo, Y., Mao, Q., Li, M., Xiao, L., Yang, D., Ban, Y., Sun, H., & Yu, P. S. (2025). Harnessing multiple large language models: A survey on LLM ensemble. arXiv preprint arXiv:2502.18036.

Christensen, D. (2007). Epistemology of disagreement: The good news. *Philosophical Review*, 116(2), 187–217.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714), eadq1814.

Dennett, D. C. (1987). *The intentional stance*. MIT Press.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR 235, 11733–11763. arXiv preprint arXiv:2305.14325.

Elga, A. (2007). Reflection and disagreement. *Noûs*, 41(3), 478–502.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681–694.

Freiman, O. (2024). AI-testimony, conversational AIs and our anthropocentric theory of testimony. *Social Epistemology*, 38(4), 476–490.

Geanakoplos, J. D., & Polemarchakis, H. M. (1982). We can't disagree forever. *Journal of Economic Theory*, 28(1), 192–200.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70, 1321–1330.

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777).

Hao, X., Wu, Z., Qiu, Y.-X., Xiao, C., Xu, R., Zheng, S., & Qin, J. (2026). Not All Flips Are Conformity: Decomposing Stance Convergence in Multi-Agent LLM Debate. *arXiv preprint arXiv:2606.00820*.

Herrmann, D. A., & Levinstein, B. A. (2025). Standards for belief representations in LLMs. *Minds and Machines*, 35:5 (2025) 1-25.

Houlsby, N., Huszár, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745.

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The platonic representation hypothesis. *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235.

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. arXiv preprint arXiv:1805.00899.

Kelly, T. (2010). Peer disagreement and higher-order evidence. In R. Feldman & T. A. Warfield (Eds.), *Disagreement* (pp. 111–174). Oxford University Press.

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., & Perez, E. (2024). Debating with more persuasive LLMs leads to more truthful answers. arXiv preprint arXiv:2402.06782.

Konigsberg, A. (2013). The problem with uniform solutions to peer disagreement. *Theoria*, 79(2), 96–126.

Lackey, J. (2008). A justificationist view of disagreement's epistemic significance. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Social Epistemology* (pp. 298–325). Oxford University Press.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30.

Li, J., Zhang, Q., Yu, Y., Fu, Q., & Ye, D. (2024). More agents is all you need. *Transactions on Machine Learning Research*.

Li, K., Patel, O., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Marks, S., & Tegmark, M. (2023). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. arXiv preprint arXiv:2310.06824.

McKinsey & Company. (2025). McKinsey & Company. (2025). *The state of AI in 2025: Agents, innovation, and transformation*. QuantumBlack, AI by McKinsey. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

Milgrom, P., & Stokey, N. (1982). Information, trade and common knowledge. *Journal of Economic Theory*, 26(1), 17–27.

Monderer, D., & Samet, D. (1989). Approximating common knowledge with common beliefs. *Games and Economic Behavior*, 1(2), 170–190.

Naik, N. (2024). Probabilistic consensus through ensemble validation: A framework for LLM reliability. arXiv preprint arXiv:2411.06535.

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs*, 36(2), 249–275.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. *The Twelfth International Conference on Learning Representations (ICLR 2024)*. https://doi.org/10.48550/arXiv.2310.13548

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Xiao, Y., Sun, E., Luo, D., & Wang, W. (2024). TradingAgents: Multi-agents LLM financial trading framework. arXiv preprint arXiv:2412.20138.

Yang, E., & Wang, D. (2026). Benchmark illusion: Disagreement among LLMs and its scientific consequences. arXiv preprint arXiv:2602.11898.

Zhao, T., Lyu, J., Jones, S., Garber, H., Pasquali, S., & Mehta, D. (2025). AlphaAgents: Large language model based multi-agents for equity portfolio constructions. arXiv preprint arXiv:2508.11152.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.

Zhu, X., Zhang, C., Stafford, T., Collier, N., & Vlachos, A. (2025). Conformity in large language models. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

[i] The assumption is surfaced in several research programs. Irving et al. (2018) proposed AI safety via debate on the premise that two AI agents arguing opposing sides of a question would surface errors and expose weaknesses in each other's reasoning, a premise that borrows its epistemic logic from adversarial proceedings between informed human advocates. Du et al. (2024) demonstrated that multi-agent debate among language models improves factuality and reasoning accuracy, framing the improvement as evidence that "diverse perspectives" among models can "correct each other's mistakes," language that implicitly attributes the epistemic properties of informed disagreement to model divergence. Khan et al. (2024) extended this line, showing that debate with more persuasive LLMs leads to more truthful answers as judged by humans, again treating the debate dynamic as epistemically productive rather than as a sampling strategy. The self-consistency approach introduced by Wang et al. (2023) treats agreement among multiple samples from the same model as a reliability signal, effectively using convergence across stochastic runs as a proxy for epistemic confidence. And the LLM-as-judge paradigm (Zheng et al., 2023) treats model evaluations of other models' outputs as carrying independent evaluative authority. In each case, the epistemic weight attributed to model agreement or disagreement borrows from the logic of human expert consensus without establishing that the conditions underwriting that logic are present.