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

Amir Konigsberg

Abstract

Aumann's Agreement Theorem (1976) establishes that two Bayesian rational agents with common priors and common knowledge of each other's posterior beliefs cannot agree to disagree. Their posteriors must coincide. This paper applies Aumann's framework to AI agents built on large language models (LLMs), a domain in which the theorem's conditions appear, at first glance, to be unusually well satisfied. LLMs trained on overlapping data are often assumed to share something like common priors, and in multi-agent protocols their outputs are shared between participants. Yet LLM-based agents routinely produce divergent outputs on identical inputs, and multi-agent systems built from them are increasingly deployed in debate, deliberation, and consensus protocols that implicitly treat this divergence as epistemically meaningful. We argue that Aumann's theorem fails to apply to these agents not because the prior or rationality conditions are violated in the familiar ways they are violated for humans, but for a more fundamental reason: LLMs do not possess beliefs in the sense the theorem requires. Their outputs are samples from conditional probability distributions over token sequences, not reports of posterior probabilities conditioned on private information. We formalize the distinction between genuine disagreement, which carries epistemic content because it signals the existence of unshared evidence [or beliefs?], and what we term *pseudo-disagreement*, which has the surface form of disagreement but arises from stochastic variation in generation processes that lack epistemic states. We show formally that pseudo-disagreement does not satisfy the informational conditions that make genuine disagreement epistemically valuable, and we trace the implications for multi-agent debate protocols, consensus methods, LLM-as-judge paradigms, and the broader practice of treating AI outputs as bearing on questions of truth. Our analysis applies specifically to autoregressive language models and the multi-agent systems built from them; AI systems with fundamentally different architectures, such as those maintaining explicit world models or calibrated Bayesian uncertainty estimates, may require separate treatment.

1. Aumann's Theorem and Its Epistemic Significance

In 1976, Robert Aumann proved a result that startled economists, philosophers, and decision theorists in equal measure: 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. 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 respectively. Each agent i observes which element of Π_i contains the true state ω, and updates via Bayes' rule to form a posterior $P(A \mid \Pi_i(\omega))$ for some event A. Aumann proved that if the agents' posteriors $q_1 = P(A \mid \Pi_1(\omega))$ and $q_2 = P(A \mid \Pi_2(\omega))$ are common knowledge, then $q_1 = q_2$.

The theorem's force lies in what it excludes. Under its conditions, persistent disagreement between rational agents is impossible. If two agents who share a prior and know each other's posteriors still disagree, then at least one of the theorem's conditions must be violated: either their priors differ, their updating is non-Bayesian, or their posteriors are not actually common knowledge.

The significance for epistemology is profound. If the theorem's conditions held in practice, disagreement would always be diagnostic. Every disagreement would signal either a difference in priors (different fundamental assumptions about the world), a difference in information (one agent knows something the other doesn't), or a failure of rationality. Conversely, when a rational agent learns that another rational agent, sharing her prior, has reached a different conclusion, she should treat this as evidence that she is missing something, because under the theorem's conditions, the disagreement cannot be sustained once both agents fully process its implications.

This epistemic function of disagreement, its capacity to signal the presence of unshared information, is what makes it valuable in human institutions. Peer review works because reviewer disagreement indicates that different experts, examining the same evidence, have reached different conclusions, prompting closer scrutiny. Deliberative democracy presupposes that disagreement among informed citizens reflects genuine differences in perspective that collective deliberation can surface and integrate. Adversarial legal proceedings assume that opposing counsel's disagreement reflects genuinely different interpretations of the evidence. In each case, the informational content of disagreement does the epistemic work.

The question this paper asks is whether AI agents built on large language models can participate in this epistemic economy: whether their "disagreements" carry the informational content that makes disagreement valuable, or whether the divergence in their outputs is a fundamentally different phenomenon masquerading in familiar epistemic clothing.

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

The conditions of Aumann's theorem, common priors, Bayesian rationality, and the possibility of communicating posteriors, seem, at first inspection, more naturally satisfied by LLM-based agents than by humans. Each condition merits examination.

2.1 Common Priors

The common prior assumption has been the most contested condition of Aumann's theorem in its application to human agents. Humans develop their beliefs through radically different life experiences, cultural contexts, and cognitive histories. The assumption that two humans share a prior probability distribution over states of the world is, at best, an idealization.

LLM-based agents trained on the same data might seem to satisfy this condition more naturally. Two instances of the same model share identical weights, and therefore identical output distributions for any given input, which is as close to a "common prior" as one could hope for in a non-Bayesian system. Even across model families (say, GPT-4 and Claude), substantial overlap in training data, since the internet is finite and the largest corpora draw from much of the same material, creates an intuition that the models share a common epistemic foundation.

But this intuition conflates training data with priors. In Bayesian terms, the prior is the full probability distribution the agent brings to bear before conditioning on evidence. For an LLM, the analog of the prior is encoded in the model's parameter weights, not in the training data. Different architectures, different random initializations, different optimization trajectories, and different fine-tuning procedures all produce different weights, and therefore different output distributions, even when the training data is identical. Two models trained on the same corpus with different random seeds will encode different probability distributions over outputs. Overlapping training data is an input to the process that produces the prior. It is not the prior itself. The common prior condition fails for any pair of distinct models, and fails trivially.

Nonetheless, the intuition persists. Researchers working on multi-agent LLM systems have implicitly relied on the assumption that models trained on similar data should, in some sense, share a common epistemic foundation, making their disagreements informative in the way that disagreements between experts with shared training are informative[i].

2.2 Consistent Inference

Bayesian rationality requires that agents apply a consistent inference rule: the same question, however presented, warrants the same answer. LLMs do not have this property. The same question, phrased differently, can produce contradictory outputs. A Bayesian agent asked whether P is true and then asked whether not-P is false must give consistent answers. An LLM may not. The forward pass is a deterministic computation, but determinism is not rationality. Any function is deterministic given fixed inputs. What LLMs lack is consistency at the level of propositions, which is the level at which Aumann's rationality condition operates.

2.3 Communication of Posteriors

Aumann's theorem requires common knowledge of posteriors: each agent knows the other's posterior, knows that the other knows their posterior, and so on. In human settings, this condition is approached only through explicit communication, and common knowledge is never literally achieved. In multi-

agent LLM systems, by contrast, agent outputs are typically fully visible to other agents. In debate protocols, each agent's response is provided in its entirety to the other agents. The communication condition seems not merely satisfied but trivially so.

Taken together, a picture emerges in which LLM-based agents are widely treated as if they satisfy Aumann's conditions: overlapping training data standing in for common priors, deterministic computation standing in for consistent inference, output sharing standing in for communication of posteriors. As the preceding subsections suggest, each of these analogies is weaker than it appears. But they have been close enough in surface form to ground an implicit premise in several active research programs. Multi-agent debate for AI alignment (Irving et al., 2018; Du et al., 2023) assumes that LLM agents arguing opposing positions will surface errors and converge toward truth. LLM-as-judge paradigms (Zheng et al., 2023) assume that language model evaluators bring something like independent judgment. Ensemble methods and multi-model consensus assume that agreement among models constitutes confirmation.

The premise is wrong. The next section explains why.

3. Why the Theorem Does Not Apply

3.1 Three Candidate Explanations

LLM-based agents produce divergent outputs. Different models give different answers to the same question. The same model gives different answers across runs. Models placed in debate settings argue, revise, and sometimes fail to converge. Three explanations present themselves, and distinguishing between them matters for everything that follows.

Explanation (a): Different priors. Models trained on different data, with different architectures, or fine-tuned with different procedures have, in effect, different priors. Their disagreement is analogous to disagreement between humans from different backgrounds: it reflects genuinely different starting assumptions, and communication could in principle resolve it. Under this explanation, AI disagreement is epistemically meaningful in something like the way human disagreement is.

Explanation (b): Bounded rationality. Models operate under computational constraints, finite context windows, approximate inference, lossy compression of training data, that prevent them from achieving the Bayesian ideal. Their disagreement is analogous to disagreement between humans who lack the cognitive resources to fully process their shared information. Under this explanation, AI disagreement is noisy but still partially informative: it reflects computational limitations rather than an absence of epistemic content.

Explanation (c): Absence of beliefs. LLMs do not have beliefs in the sense Aumann's theorem requires. They have output distributions: conditional probability distributions over token sequences given an input. These distributions are not posteriors in the Bayesian sense. They are not the result of conditioning a prior on privately observed evidence. They are the result of a deterministic function (the forward pass) applied to an input and then sampled stochastically. The "disagreement" between

two LLM-based agents is not a disagreement between beliefs; it is a divergence between samples drawn from stochastic processes.

We argue that while (a) and (b) capture real phenomena, they are subordinate to (c), which identifies the fundamental reason Aumann's framework does not apply. The next two subsections make this argument precise.

3.2 The Belief Condition

Aumann's theorem is a theorem about beliefs, formalized as probability assignments that an agent maintains about the world and that are updated in light of evidence. Beliefs, in this framework, have several properties that LLM outputs lack.

Beliefs are about states of the world. An agent's posterior $P(A \mid e)$ is a probability assigned to an event A in light of evidence e. The belief is directed at A: it is about whether A obtains. An LLM's output distribution $P(t_1, t_2, ..., t_n \mid \text{prompt})$ is a probability distribution over token sequences. It is not, in any direct sense, a probability assigned to a proposition about the world. It is a probability assigned to a sequence of symbols, calibrated by how frequently similar sequences appeared in similar contexts in training data.

The distinction matters formally. Aumann's theorem operates in a space of propositions about states of the world. The "probabilities" that LLMs assign operate in a space of token sequences. The mapping between these two spaces is not straightforward. A model may produce "The probability of rain tomorrow is 70%" with high probability, not because it has assigned 0.7 to the proposition that it will rain, but because that string is a probable continuation of the prompt in its training distribution. The surface form resembles a belief report. The underlying process does not involve belief formation.

Beliefs are updated by evidence. Bayesian agents revise their beliefs when they observe new evidence. The posterior is the prior conditioned on the observation. LLMs do not update in this sense. They process a fixed context window and produce an output. They do not maintain persistent belief states across interactions (absent external memory systems). Each generation is a fresh computation, not a revision of a prior belief in light of new evidence. When a model "changes its mind" during a multi-turn debate, what has changed is the input (the context now includes the other agent's response), not an internal belief state. The model is producing a new output conditioned on a new input, not updating a prior in light of evidence. The distinction is between *stateless conditional generation* and *belief revision*. They look similar in their outputs. They are different in kind.

Beliefs are held by a subject. This may seem like a philosophical quibble, but it matters for the formal structure. Aumann's agents are entities *for whom* the propositions in question are meaningful: agents who have stakes in the outcomes, who act on their beliefs, whose beliefs cohere (or fail to cohere) into a worldview. The theorem's power derives from the assumption that each agent's posterior reflects a coherent assessment of the evidence available to them. An output distribution does not have this property. It is a mathematical object, not a stance.

3.3 The Information Condition

Even if we set aside the question of whether LLM-based agents have beliefs, Aumann's theorem requires that disagreement arise from differential access to information. Agent 1 observes that the true state ω falls in partition element $\Pi_1(\omega)$; agent 2 observes that it falls in $\Pi_2(\omega)$. Their different posteriors reflect their different information, and communication of posteriors allows each to infer something about the other's private observation.

LLM-based agents do not have private information in this sense. Two instances of the same model, given the same prompt, have access to exactly the same "information" (the prompt and their shared training data). Their divergent outputs are not the result of different observations. They are the result of stochastic sampling from the same (or very similar) output distributions.

When two different models disagree, the divergence reflects differences in training data, architecture, and fine-tuning. But these differences are not "private information" in Aumann's sense. Private information, in the theorem, is information about the state of the world that one agent has and the other lacks. Different training data is a difference in the function that maps inputs to outputs, not a difference in observations about the world. The models have not observed different evidence about the proposition in question; they have been constructed by different processes that happen to produce different outputs.

This distinction, between *informational divergence* (different evidence about the world) and *functional divergence* (different input-output mappings), is the crux of the argument. Aumann's theorem tells us something important about informational divergence: it can be resolved through communication of posteriors, and in the interim it signals the presence of unshared evidence. The theorem tells us nothing about functional divergence, because functional divergence is not within its scope.

4. A Formal Model of Pseudo-Disagreement

4.1 Definitions

We formalize the distinction between genuine and pseudo-disagreement.

Definition 1 (Genuine Disagreement). Let agents 1 and 2 share a common prior P over a state space Ω. Let Π_1 and Π_2 be their respective information partitions. Agents genuinely disagree about event A at state ω if:

$$P(A \mid \Pi_1(\omega)) \neq P(A \mid \Pi_2(\omega))$$

Genuine disagreement arises from differential conditioning on a shared prior. It is informative because each agent's posterior reflects evidence the other has not observed.

Definition 2 (Pseudo-Disagreement). Let f_1 and f_2 be two stochastic generation functions mapping inputs to output distributions. Agents pseudo-disagree about input x if:

$$f_1(x) \neq f_2(x)$$

where inequality indicates that the output distributions differ, or that distinct samples have been drawn from the same or similar distributions.

Pseudo-disagreement arises not from differential access to information about the world but from divergence in the functions themselves (different models) or from stochastic sampling (different runs of the same model). The divergence does not signal the existence of unshared evidence. It signals a difference in the generative process.

Definition 3 (Intra-Model Pseudo-Disagreement). Even a single model f can pseudo-disagree with itself. Let s_1 and s_2 be two samples drawn from $f(x)$ with different random seeds. If $s_1 \neq s_2$, the model is in intra-model pseudo-disagreement. This case makes the distinction with genuine disagreement especially clear: nothing about the state of the world differs between the two samples. The divergence is pure aleatoric noise.

4.2 The Informational Content of Disagreement

We now state the key result informally (a fully formal treatment would require a measure-theoretic framework beyond the scope of this paper, but we indicate the structure of the argument).

Proposition (Epistemic Content of Genuine Disagreement). Under Aumann's conditions, if agent 1 learns that agent 2's posterior for event A is $q_2 \neq q_1$, then agent 1 can infer that the true state ω lies in some subset of states consistent with agent 2 assigning posterior q_2, which generically excludes some states consistent with agent 1's current information. The disagreement is evidence. It narrows the set of possible states.

This is why Aumann's result matters epistemically: disagreement between Bayesian agents with common priors is *never just disagreement*. It is always a signal of evidence that the other agent possesses and you do not. Rational agents who take this seriously will revise their beliefs, which is the mechanism through which the theorem guarantees convergence.

Proposition (Epistemic Vacuity of Pseudo-Disagreement). When two LLM-based agents pseudo-disagree about input x, the fact of their divergence does not narrow the set of possible states of the world. The divergence is attributable to (a) differences in the generative functions f_1 and f_2 (different training, architecture, fine-tuning) or (b) stochastic sampling. Neither (a) nor (b) constitutes evidence about the proposition expressed by x. Learning that model 2 produced a different output does not give model 1 (or a human observer) grounds for updating beliefs about the world. It gives them grounds for updating beliefs about model 2's output distribution, which is a fact about the model, not a fact about the world.

The asymmetry is stark. In genuine disagreement, the disagreement *is about* the state of the world, and resolving it can improve both agents' epistemic positions. In pseudo-disagreement, the disagreement is about the outputs of generative processes, and resolving it (by, say, averaging the outputs or selecting

the majority response) does not constitute epistemic progress in the sense Aumann's framework describes.

4.3 The Aleatoric-Epistemic Distinction

The analysis maps onto a well-known distinction in statistics and machine learning between two types of uncertainty:

Epistemic uncertainty arises from ignorance: it is reducible by acquiring more information. When a doctor is uncertain about a diagnosis because she hasn't yet seen the lab results, her uncertainty is epistemic. The lab results exist; obtaining them will reduce her uncertainty.

Aleatoric uncertainty arises from inherent randomness: it is irreducible by information. When a physicist is uncertain about where a photon will land, this uncertainty reflects the stochastic nature of the process itself, not ignorance of hidden variables (under standard quantum mechanics).

Genuine disagreement between Bayesian agents reflects epistemic uncertainty: each agent's posterior differs because they have different evidence, and sharing that evidence would reduce the disagreement. LLM pseudo-disagreement reflects primarily aleatoric uncertainty: the divergence in outputs arises from stochastic sampling in the generation process, or from the effectively random differences in how different training runs settled on different parameter configurations.

This mapping clarifies what is lost when pseudo-disagreement is treated as genuine disagreement. Protocols designed to harness disagreement, debate, deliberation, peer review, derive their epistemic value from the assumption that disagreement signals unshared evidence (epistemic uncertainty). When the disagreement actually reflects stochastic variation (aleatoric uncertainty), these protocols run on empty. They go through the motions of epistemic negotiation without the epistemic substance.

5. Implications

5.1 Multi-Agent Debate for Alignment

Multi-agent debate has been proposed as a mechanism for improving AI safety and alignment (Irving et al., 2018; Du et al., 2023; Khan et al., 2024). The basic idea is that LLM agents arguing opposing sides of a question will expose each other's errors, and a human judge can identify the stronger argument. The approach draws, implicitly but unmistakably, on the epistemic logic of adversarial proceedings: truth emerges from the clash of opposing positions.

Our analysis suggests this analogy is structurally flawed. In an adversarial legal proceeding, opposing counsel disagree because they have genuinely different interpretations of the evidence, and their disagreement is informative because each interpretation reflects a perspective the other may have missed. In multi-agent LLM debate, the agents "disagree" because they have been assigned opposing positions and generate token sequences that argue for those positions. The disagreement is not the

product of different evidence or different rational assessments of shared evidence. It is the product of an instruction to generate text arguing for position X vs. position Y.

This does not mean multi-agent debate is useless. It may surface considerations that a single agent would not produce (by sampling from different regions of the output distribution). It may improve the persuasiveness or completeness of the arguments presented to the human judge. But its epistemic status is different from what the adversarial analogy suggests. The debate does not harness the informational content of genuine disagreement. It harnesses the diversity of a stochastic generation process, which is a different and weaker epistemic resource.

5.2 LLM-as-Judge and Consensus Methods

The use of language models as evaluators of other models' outputs (Zheng et al., 2023) and the practice of taking the consensus of multiple models as a reliability signal both assume that LLM agreement and disagreement carry epistemic weight. If three models agree on an answer, this is treated as confirmation. If they disagree, this is treated as a signal of uncertainty.

Under our framework, this practice conflates two very different phenomena. When three human experts agree on a diagnosis, their agreement is informative because each brought independent judgment, training, and possibly different evidence to bear. Their convergence raises the probability that the diagnosis is correct, because each expert's agreement represents an independent (or at least partially independent) epistemic endorsement.

When three language models agree on an output, the situation is fundamentally different. Models trained on overlapping data with similar architectures will tend to produce similar outputs not because each has independently assessed the evidence, but because each is sampling from a similar region of output space. Their agreement reflects the shared structure of their training distributions, not independent epistemic endorsement. The "confirmation" is an artifact of shared construction, not converging evidence.

Disagreement among models is correspondingly less informative. When two LLMs disagree, this signals a divergence in their output distributions, which may reflect differences in training data, architecture, or sampling. It does not signal that one model has evidence the other lacks, because LLMs do not have evidence in the relevant sense.

5.3 The Epistemic Peer Problem

Recent philosophical work on peer disagreement has explored what a rational agent should do when confronted with disagreement from an epistemic peer, someone equally competent and equally well-informed (Christensen, 2007; Elga, 2007; Kelly, 2010). The standard positions range from "conciliationism" (you should move your credence toward the peer's view) to "steadfastness" (you can maintain your view if you have good reason). The debate presupposes that the peer has genuine beliefs formed through rational assessment of evidence.

The practice of treating an LLM as an epistemic peer, asking it for a "second opinion," deferring to it when it disagrees with your initial judgment, relying on it to check your reasoning, imports the assumptions of the peer disagreement literature without the conditions that make those assumptions warranted. When a human expert disagrees with you, the disagreement is potentially evidence that you have overlooked something: the expert has assessed the same body of evidence and reached a different conclusion, which under Aumann-like conditions should prompt you to revise your view. When an LLM disagrees with you, the disagreement reflects the output of a generative process that has no beliefs about the matter, has not assessed evidence, and has not reached a conclusion. The psychological force of the disagreement (which, as work on AI persuasion demonstrates, can be substantial) is unmoored from the epistemic content that would justify treating it as informative.

This has direct implications for the erosion of independent judgment. If humans treat LLM disagreement as carrying the same epistemic weight as human peer disagreement, they will revise their beliefs in response to signals that carry no information about the truth of the matter. Over time, this produces a systematic distortion: beliefs shaped not by evidence and rational deliberation but by the output tendencies of generative models, which reflect training distributions, fine-tuning objectives, and the stochastic properties of sampling, none of which are epistemically relevant to the questions at hand.

5.4 When LLM Disagreement Is Useful (and Why)

We do not claim that divergence among LLM outputs is never useful. It can be useful in at least two ways, but neither involves the epistemic mechanism that genuine disagreement employs.

First, sampling multiple outputs from a model (or from multiple models) can increase coverage of the space of possible responses. If the correct answer is a low-probability output under any single sample but a moderate-probability region of the output distribution, multiple samples increase the chance of finding it. This is a search strategy, not an epistemic one. It is closer to a Monte Carlo method than to a deliberation.

Second, divergence among outputs can serve as a proxy for the difficulty or ambiguity of the input. If many models produce the same answer, the input is probably unambiguous; if they diverge widely, the input may be genuinely difficult or poorly specified. This is a useful diagnostic, but it is a diagnostic about the relationship between the input and the models' training distributions, not about the state of the world. High divergence tells you that the models are uncertain in the aleatoric sense, not that they possess conflicting evidence in the epistemic sense.

Recognizing these genuinely useful functions while clearly distinguishing them from the epistemic functions of genuine disagreement would improve both the design and the interpretation of multi-agent LLM systems.

6. Related Work

The application of formal epistemology to AI systems is a growing field but remains underdeveloped relative to its importance. Relevant work falls into several categories.

On the formal side, Aumann's original result (1976) and its extensions (Geanakoplos & Polemarchakis, 1982; Milgrom & Stokey, 1982) established the theoretical foundations. The impossibility of agreeing to disagree under common priors has been extended to dynamic settings, approximate agreement, and bounded rationality (Monderer & Samet, 1989; Aaronson, 2005). To our knowledge, no prior work has applied this framework to the specific case of LLM-based agents, though the question of whether AI systems can be modeled as Bayesian agents has been discussed informally in the alignment literature.

On the multi-agent AI side, Irving et al. (2018) proposed AI safety via debate, in which two LLM agents argue before a human judge. Du et al. (2023) demonstrated that multi-agent debate among language models can improve reasoning accuracy. Khan et al. (2024) explored debate as a scalable oversight mechanism. These papers evaluate the empirical effectiveness of debate protocols but do not examine the epistemic foundations that would justify interpreting the debate as epistemically productive in the sense formal epistemology requires.

On the philosophical side, the peer disagreement literature (Christensen, 2007; Elga, 2007; Kelly, 2010; Lackey, 2010) has developed sophisticated accounts of how rational agents should respond to disagreement, but has not addressed the case of artificial agents that lack beliefs. Recent work on AI testimony (Freiman, 2024) and machine epistemology (Floridi & Chiriatti, 2020) has begun to explore whether AI outputs can function as testimony or evidence, but has not connected this question to the formal framework of agreement and disagreement.

Our contribution bridges these literatures. By applying Aumann's framework to LLM-based agents and showing precisely where and why it fails, we provide a formal basis for claims that have so far been made only informally: that LLM outputs lack the epistemic status of genuine beliefs, and that treating them as if they had this status leads to specific, identifiable errors in system design and epistemic practice.

7. Objections and Responses

"Models do have something like beliefs. Internal representations track truth." Recent interpretability work has shown that LLMs develop internal representations that correlate with the truth values of propositions (Li et al., 2023; Marks & Tegmark, 2024). This is a genuine finding. But truth-tracking representations are not the same as beliefs. A thermometer tracks temperature, but it doesn't believe it's cold. The relevant question is whether the model's outputs are *reports of* these internal representations in the way that a human's verbal assertion is a report of their beliefs. The evidence suggests they are not: the relationship between internal representations and outputs is mediated by the generation process, fine-tuning, and RLHF, all of which can and do decouple what the model "represents" internally from what it says. Unfaithful chain-of-thought is one well-documented instance of this decoupling (Turpin et al., 2024).

"Disagreement between differently trained models does carry information." We agree that inter-model divergence can carry information, specifically, information about which regions of output space different training processes converge on. This information can be useful for model development and evaluation. But it is information about the models, not information about the world. The distinction matters. Treating inter-model divergence as evidence about the world, rather than evidence about the models' training distributions, is the category error our analysis identifies.

"The common prior assumption already fails for humans, so the theorem doesn't apply to humans either." Correct, strictly speaking. The theorem's conditions are idealized. But the epistemic logic the theorem formalizes, that disagreement between agents with shared foundations signals unshared information, holds approximately for humans to the degree that the conditions are approximately satisfied. Our argument is not that the theorem applies perfectly to humans and fails for LLMs. It is that the specific condition that fails for LLMs (the belief condition) is different in kind from the conditions that fail for humans (approximate common priors, bounded rationality), and this difference has consequences for how LLM disagreement should be interpreted.

"You're setting an impossibly high bar. If AI outputs are never epistemically relevant, they're useless." We do not claim they are never useful, as Section 5.4 specifies. We claim they are not epistemically relevant in the specific sense that genuine beliefs are: they do not carry the informational content that warrants the epistemic practices (peer disagreement reasoning, adversarial deliberation, consensus as confirmation) that are currently being applied to them. Recognizing this distinction does not make AI systems less useful. It makes their use more precisely calibrated to what they actually provide.

8. Conclusion

Aumann's Agreement Theorem tells us something deep about the nature of rational disagreement: when it occurs between agents who share a common epistemic foundation, it is always informative, always a signal that one agent possesses evidence the other lacks, always resolvable in principle through full communication. This informational content is what makes disagreement epistemically valuable, and it is what justifies the institutions, practices, and protocols that harness disagreement in the service of truth.

LLM-based agents do not participate in this epistemic economy. They produce divergent outputs, and these divergences have the surface form of disagreement: two agents, same question, different answers. But the divergence arises from a fundamentally different source than genuine disagreement. It arises from stochastic variation in generative processes that lack beliefs, have not observed evidence, and do not maintain epistemic states. What LLM-based agents do is not agree to disagree. What they do is generate different samples from a process that has no opinion on the matter.

The implications extend beyond the formal framework. Multi-agent debate, consensus methods, LLM-as-judge paradigms, and the everyday practice of treating a language model as an epistemic peer all draw, implicitly, on the assumption that AI agreement and disagreement carry the same epistemic

weight as their human counterparts. Our analysis shows that this assumption is unwarranted. The weight is not there. The form is the same, but the substance is absent.

This does not mean these systems should be abandoned. It means they should be understood for what they are: methods for sampling diverse outputs from generative processes, not methods for harnessing the epistemic content of genuine disagreement. Calibrating our expectations to this reality, rather than to the epistemic analogy that currently governs design and interpretation, is a precondition for using these systems wisely.

References

Note: All citations require independent verification before submission.

Aaronson, S. (2005). The complexity of agreement. *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, 634–643.

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

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

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. *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. *Philosophy and Technology*, 37(1), 1–22.

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

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., Raber, A., ... & Perez, E. (2024). Debating with more persuasive LLMs leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Lackey, J. (2010). A justificationist view of disagreement's epistemic significance. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Social Epistemology* (pp. 298–325). Oxford University Press.

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. (2024). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

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.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2024). 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., ... & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *Proceedings of the International Conference on Learning Representations*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.

---

[i] The assumption surfaces in several distinct 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. (2023) 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.