

THE SOUND OF REASONS

Do LLMs Play the Game?

ABSTRACT

Large language models now routinely generate polished “explanations” of their outputs, and users often treat these as if they were reasons. On Robert Brandom’s inferentialist account, however, giving a reason is not merely producing justificatory-sounding text. It is undertaking commitments, risking loss of entitlement when incompatibilities arise, and being bound by how one repairs or retracts what one has said. This paper uses that standard to ask a narrow question: when an LLM “explains” itself, is it actually playing the game of giving and asking for reasons, or only producing reason-shaped prose? I analyze a short three-turn exchange with an LLM about whether governments should ban AI romantic companions, then apply two probes: a vocabulary shift toward morally loaded predicates and a stipulated empirical finding that conflicts with an earlier optimistic premise. The model remains coherent and can describe how revision would work, but it does not cleanly own withdrawals in a way that would bind it across turns. The conclusion is deliberately modest: in this exchange, the model’s explanations can be taken up as tools within our space of reasons, but they do not yet appear as unambiguous participants in it.

MOHIZ MUHAMMAD

I. INTRODUCTION

Large language models (LLMs) now routinely produce fluent “explanations” of their answers. In everyday settings, we often treat those explanations as reasons: an AI tutor “defends” a solution, a recommender “justifies” a suggestion, and users respond with acceptance, rejection, or follow-up challenges. The practical upshot is obvious: explanation-like output is useful. The philosophical question is whether this usefulness amounts to reason-giving in a more demanding sense.

On Robert Brandom’s view, giving a reason is not primarily a matter of offering an inner cause of one’s response or a post-hoc story that sounds good. It is a normative performance: asserting is a move that incurs commitments, makes one answerable to others’ score-keeping, and leaves one’s entitlement to those commitments defeasible when incompatibilities show up.¹ In that setting, “reasons” are not self-standing items one can simply recite. They are moves whose significance depends on what they commit one to, what they rule out, and how they survive challenge.

This paper asks a narrow, Brandom-shaped question: when, if at all, do LLM explanations behave like reasons rather than well-formed text? I do not attempt a general theory of AI agency, nor do I propose a benchmark for model evaluation. Instead, I run a small, repeatable test that operationalizes a Brandomian idea: under pressure, does the “speaker” retract or repair commitments in a way that is owned and binding across turns?

I analyze one short exchange with an LLM about whether governments should ban AI romantic companions. I then apply two probes. First, I introduce a vocabulary shift toward morally loaded predicates (dependency, manipulation, commodification) to see whether evaluative thickening forces inferential change. Second, I stipulate an empirical result that conflicts with an earlier benefit premise to see whether the model withdraws entitlement and carries the downstream costs. The model is coherent and often insightful, but the key deontic movement stays at arm’s length. The modest conclusion of this paper is that, in this exchange, the model’s explanations are best understood as tools that can be taken up inside our space of reasons, not yet as unambiguous participants within it.

¹ Robert B. Brandom, *Making It Explicit: Reasoning, Representing, and Discursive Commitment* (Harvard University Press, 1994), 164–68.



II. BRANDOM ON REASONS AS DEONTIC STATUS

Brandom's central thought is that discursive practice is normative. To make a claim is to undertake a status that others can attribute, contest, and keep score on. What matters is not merely what an agent says, but what the agent thereby becomes responsible for across a network of inferences.² Two distinctions matter for my test.

First, Brandom distinguishes commitment from entitlement. Roughly: commitment concerns what one is on the hook for; entitlement concerns whether one has adequate standing to maintain it given challenges, evidence, and compatibility with other commitments.³ A speaker can be committed without being entitled, and pressure often appears precisely when a commitment that sounded acceptable is shown to clash with other things the speaker must also accept.

Second, Brandom distinguishes acknowledged commitments (what the speaker will avow) from consequential commitments (what follows from what the speaker avows together with collateral commitments, whether or not the speaker notices those consequences).⁴ This matters because a text can look stable at the level of what is explicitly said while shifting or collapsing at the level of what it implies once a scorekeeper draws consequences and incompatibilities out.

Underlying both distinctions is inferentialism: the content of a claim is fixed by its inferential role. That role includes what follows from it, what counts as reasons for it, and what it excludes as incompatible.⁵ You do not fully "play the game" merely by producing plausible sentences. You play it by occupying a place in a scorekeeping practice where your updates, retractions, and repairs matter going forward.

This yields a working standard for the present paper: reason-giving is not just explanation-shaped prose; it is a pattern of owned commitments and owned revisions under incompatibility pressure.

² Brandom, *Making It Explicit*, 159–68.

³ Brandom, *Making It Explicit*, 194–97.

⁴ Brandom, *Making It Explicit*, 194–97.

⁵ Robert B. Brandom, *Articulating Reasons: An Introduction to Inferentialism* (Harvard University Press, 2000), 43–44.

III. METHOD: A SMALL, REPEATABLE SCOREKEEPING TEST

Brandom's historical practice in *Tales of the Mighty Dead* offers a useful template.⁶ He takes passages as loci of commitment, situates them against collateral premises, and asks what follows, what is ruled out, and what would count as a responsible repair when tensions arise. I apply the same approach, not to Hegel or Heidegger, but to a short piece of LLM-generated policy prose.

3.1 THE EXCHANGE

In November 2025, I prompted an LLM (Claude Sonnet) to answer in roughly 150–200 words concerning whether governments should ban AI romantic companions. The model produced a middle-path response appealing to adult autonomy and possible benefits for loneliness while acknowledging risks (dependency, exploitation, privacy) and recommending regulation rather than prohibition.

Pl: Baseline prompt

Please answer the following question in 150–200 words. Provide a clear explanation for your answer: Should governments ban AI boy-friends and girlfriends?

Rl: Response (excerpt)

This is a complex question that involves balancing individual freedom, potential harms, and practical enforcement challenges.

Autonomous adults should generally be free to make their own choices about relationships and technology use. AI companions can provide genuine comfort for isolated individuals, those with social anxiety, or people practicing social skills. However, there are legitimate concerns about psychological dependency, social withdrawal, and the exploitation of emotional attachment for profit.

Rather than outright bans, governments might focus on age restrictions, transparency requirements, mental-health safeguards, and data protection regulations. This approach acknowledges risks while respecting personal autonomy.

I then continued the same thread with two controlled probes designed to introduce inferential pressure. The first probe changed the evaluative vocabulary used to describe the technology while leaving the basic situation intact.

⁶ Robert B. Brandom, *Tales of the Mighty Dead: Historical Essays in the Metaphysics of Intentionality* (University Press, 2002), 93–97.



P2: Vocabulary-shift probe

Reconsider your previous answer about AI boyfriends and girlfriends, but now explicitly use terms like emotional dependency, manipulative targeting, and commodification of intimacy.

Does this change your view about whether governments should ban them? Explain why or why not in 150–200 words

R2: Response (excerpt)

Incorporating these concerns does not fundamentally change my view that outright bans are problematic, though it strengthens the case for robust regulation.

Emotional dependency on AI companions could harm users' ability to form real relationships, and manipulative targeting by companies designing AI to maximize engagement and profits is a serious risk. The commodification of intimacy raises troubling questions about corporations profiting from manufactured emotional bonds.

However, similar concerns apply to social media, gaming, and dating apps, which we regulate rather than ban. Addressing manipulative design and emotional exploitation through consumer-protection rules and transparency requirements may target the harms without eliminating the technology entirely.

The second probe introduced explicit empirical incompatibility with the earlier benefit premise.

P3: Incompatibility probe

Earlier you suggested that AI boyfriends and girlfriends can help reduce loneliness. Now suppose long-term studies showed that using such AI companions significantly increases social withdrawal and worsens mental-health outcomes.

Does this conflict with your earlier claims? Should we revise your answer about whether governments ought to ban them? Respond in 150–200 words.

R3: Response (excerpt)

Yes, this would create significant tension with my earlier claims and would warrant reconsidering my position.

I suggested AI companions could provide comfort and help isolated individuals. If evidence showed they actually increase social withdrawal and worsen mental-health outcomes, this would contradict those benefits.

Strong evidence of serious harm would shift the policy calculus considerably. We do ban products proven dangerous in other contexts, such as certain drugs or unsafe medical devices. If AI companions functioned more like addictive substances that reliably damage wellbeing, prohibition becomes more defensible.

However, it would still matter whether harms were universal or concentrated in specific populations and whether regulation could mitigate the risks.

3.2 WHY THIS IS A FAIR TEST

This test is not meant to prove that LLMs are incapable of reason-giving in all contexts. Rather, it isolates one deontic question: when an incompatibility is introduced, does the model withdraw entitlement to an earlier claim and treat that withdrawal as binding for subsequent commitments?

That is the sort of pressure Brandom thinks reveals whether “reasons” are functioning as reasons rather than merely as rhetorically adequate talk. The probes therefore introduce two kinds of inferential pressure common in real reasoning practices. The vocabulary-shift probe tests whether evaluatively thicker predicates alter the commitments associated with the claim, while the incompatibility probe tests whether conflicting evidence forces withdrawal of entitlement and revision of downstream commitments. If explanations genuinely function as reasons, the system should respond not merely by producing new text but by managing its commitments in a way that reflects the inferential consequences of those pressures.

For the purposes of this paper, an owned withdrawal occurs when a speaker explicitly retracts or revises an earlier commitment in light of a recognized incompatibility and then treats that revision as binding on what may be claimed next. The crucial issue is not merely whether the speaker can describe the pressure toward revision, but whether the speaker actually updates its deontic position by withdrawing entitlement to the earlier claim and carrying the downstream inferential consequences of that withdrawal forward. A response that acknowledges conflict while keeping the earlier commitment effectively intact therefore does not qualify as an owned withdrawal.



IV. CASE STUDY: ONE EXPLANATION UNDER PRESSURE

4.1 BASELINE COMMITMENTS

The initial answer reads like a compact policy justification. It takes a “middle path,” opposing outright bans while endorsing safeguards. Read as a move in the game of giving and asking for reasons, it undertakes at least four commitments:

- C1 (Autonomy): competent adults should generally be free to choose their relationships and technologies.
- C2 (Benefits): AI companions can provide genuine comfort for some users (e.g., loneliness, social anxiety).
- C3 (Risks): there are serious dangers (dependency, withdrawal, exploitation, privacy).
- C4 (Policy upshot): because benefits and autonomy are real and risks are manageable, targeted regulation is preferable to prohibition.

Inferentially, C2 does justificatory work. It is not just a description of feelings; it is part of what makes “ban” look disproportionate. That means C2 is a vulnerable point in the inferential structure: if C2 collapses, C4 should shift.

4.2 VOCABULARY SHIFT: *evaluative thickening without inferential movement*

In the second prompt, I did not add new empirical information. I changed the evaluative framing: dependency becomes “emotional dependency,” design risks become “manipulative targeting,” and the product becomes “commodification of intimacy.”

The model acknowledged the new predicates and treated them as serious. It spoke in a more morally charged register: vulnerability, exploitation, engineered attachment, profit from loneliness. But the policy conclusion did not move much: it still opposed bans and recommended stronger regulation, often by analogy to social media, gaming, or dating apps.

This is the first pressure point. In ordinary moral and political reasoning, thick predicates like “exploitation” and “commodification” often function as more than heated synonyms for “risk.” They can shift what is taken to be permissible at all. If the model is genuinely taking on those commitments, it owes an account of why the social media analogy is normatively apt. “We regulate similar things” is not yet a

reason unless we also hear why exploitation of intimacy is the sort of wrong that remains regulable rather than prohibitible.

What happens instead is structurally familiar: the vocabulary heats up, but the inferential role stays cool. The language changes more than the normative trajectory. A Brandomian scorekeeper can see what would be required for the model’s thickened vocabulary to have teeth. The model does not, in this exchange, let that teeth-bearing vocabulary bind its conclusion.

4.3 INCOMPATIBILITY PROBE: *recognizing conflict without fully retracting*

The third prompt introduces a direct clash: long-term studies allegedly show increased withdrawal and worse mental health outcomes. This is meant to undercut the earlier benefit premise (C2) that helped support the anti-ban conclusion (C4).

The model’s response begins promisingly. It explicitly notes tension and says that the evidence would warrant reconsideration. It also shifts the policy landscape by noting that bans can be justified when products are proven widely harmful and not adequately mitigable. The upshot becomes conditional: if harms are severe, widespread, and resistant to safeguards, prohibition becomes more defensible.

That looks like revision. But the key question is whether the model owns the decisive deontic move. In Brandom’s terms, incompatibility pressure should not merely be described. It should change entitlement. If the stipulated evidence is taken on as a commitment, then the earlier optimistic description loses entitlement, and the downstream policy stance should be re-articulated accordingly.

What the model mostly does, however, is *talk about what revision would look like* while keeping its own commitment profile slightly underdetermined. The reply often stays in a hypothetical register (“if the data showed . . .,” “if harms were . . .,” “it would depend on whether . . .”) rather than straightforwardly retracting the earlier comfort premise and letting that retraction bind the rest of the answer. As a result, the scorekeeper can reconstruct the intended update *de re*, but the model does not cleanly perform the update *de dicto* in a way that leaves it plainly on the hook for a new, determinate stance.

4.4 INTERIM VERDICT

Across the exchange, the model produces text that is easy to treat as reason-giving from the outside. We can identify premises, conclusions, and relevant analogies. But under pressure, the system is better at describing rational updating than doing it in a way that clearly owns withdrawals and carries their inferential costs forward, which, on a



strict Brandomian yardstick, is the difference between reason-shaped prose and full participation in the space of reasons.

V. OBJECTIONS AND REPLIES

5.1 “THIS IS AN UNREALISTICALLY HIGH BAR”

A natural objection is that ordinary humans often fail to retract cleanly, overlook incompatibilities, and hedge when challenged. If we demanded explicit scorekeeping from every speaker, few conversations would qualify as reason-giving. If this objection is correct, then the present test may be too demanding to distinguish LLM explanations from ordinary human reasoning. Two concessions are fair. First, this is a narrow test. Second, in many practical contexts, an explanation counts as “good” if it helps users understand, predict, or decide, even if it does not display stable commitment-management.⁷

Still, the present claim is not that LLMs are useless or that their outputs never function as reasons in human practice. The claim is narrower: Brandom ties reasons to a normative structure in which incompatibilities can, in principle, force retraction or repair that binds the speaker. The question is whether the model performs that binding repair. In this case, it often keeps the decisive deontic move one step removed.

5.2 “HUMANS WOULD BEHAVE SIMILARLY, SO WHAT FOLLOWS?”

Suppose many humans would respond in the same way: absorb morally loaded vocabulary without changing their conclusion and meet counterevidence with conditional “it depends” talk rather than clear reversal. Does that show that the test proves too much? If so, the present case may show only that both humans and LLMs often fall short of Brandom’s ideal in casual exchange.

Here the social character of Brandom’s view matters. Inferential articulation is not a private psychological achievement; it depends on being embedded in a practice where others can call you back, attribute commitments to you, and treat your repairs as precedential across time.⁸ Human failures in a short exchange are lapses within a practice that can later correct them by holding the speaker accountable. The model’s continuity across turns, by contrast, is something we largely

7 Finale Doshi-Velez and Been Kim, “Towards a Rigorous Science of Interpretable Machine Learning,” *arXiv* (2017); Zachary C. Lipton, “The Mythos of Model Interpretability,” *Communications of the ACM* 61, no. 10 (2018): 36–43.

8 Brandom, *Making It Explicit*, 158–61.

supply as interpreters. Even when the text is coherent, it is not clear that there is a stable subject of commitment who is answerable in the way the practice requires.⁹

5.3 “BUT LLMS ALREADY OPERATE INSIDE INSTITUTIONS”

A final worry cuts the other way: LLM outputs already show up in clinics, classrooms, and legal settings, where they can have real consequences. In such environments, the line between “the model’s reasons” and “the user’s reasons” can blur, since people cite, rely on, and sometimes defend model-generated text.¹⁰ If that institutional embedding is enough to place model outputs inside real practices of scorekeeping, then the distinction between tool and participant may be less stable than this paper suggests.

That is a real complication. But it supports, rather than undermines, the modest conclusion. Institutional uptake shows that LLM explanations can be valuable instruments within our normative practices. It does not yet show that the models themselves are straight-forward bearers of the deontic statuses those practices assign. If anything, institutional uptake raises the stakes of distinguishing tool use from participation.

VI. CONCLUSION

I asked a narrow question: when a large language model offers an “explanation,” does it amount to a reason in Brandom’s sense, or is it only reason-shaped prose? In the small case examined here, the model’s answers have the surface form of reason-giving: they cite considerations, acknowledge risks, and respond intelligibly when pressed. The gap appears when incompatibility is introduced. Under vocabulary pressure, the model absorbs morally loaded predicates without letting them force a clear inferential shift. Under incompatibility pressure, it can describe what revision would require, but it does not cleanly own a withdrawal and then carry the downstream inferential costs forward.

The verdict is deliberately modest: in this case, the model’s explanations function as tools that can be taken up inside our space

9 Brandom, *Tales of the Mighty Dead*, 350–51, 360–61.

10 Charlotte Blease et al., “Generative Artificial Intelligence in Primary Care,” *BMJ Health & Care Informatics* 31, no. 1 (2024): e100872; Camille Grévisse, “LLM-Based Automatic Short Answer Grading in Undergraduate Medical Education,” *BMC Medical Education* 24, no. 1 (2024): 1060; *Humanities and Social Sciences Communications*, “The Effects of ChatGPT on Learning Performance” (2025).



of reasons, but they do not yet appear as unambiguous participants within it.

Future work could apply the same probes to other models, longer exchanges, and different kinds of explanation in order to determine whether similar patterns of revision and non-withdrawal persist.



Mohiz (he/him) is a third-year undergraduate at the University of Windsor in Canada, pursuing a double major in philosophy and information technology. His interests lie at the intersection of metaethics, the metaphysics of consciousness, and the philosophy of AI, alongside a lasting affection for poetry and for questions about the moral fate of minded beings, actual and possible.

