

## From Reasons to Reason-Shaped Prose

Correctness depends on score.

– David Lewis, *Scorekeeping in a Language Game*

*Large language models now routinely generate polished “explanations” of their outputs, and in practice we often treat these as if they were reasons. On Brandom’s picture, giving a reason is taking up a deontic status. You undertake commitments, you can lose entitlement when incompatibilities appear, and you are bound by how you repair or withdraw what you have said. I use that standard to ask a narrow question: when an LLM “explains” itself, is it actually playing that game, or only producing reason-shaped prose? I take one short answer from Claude Sonnet to the prompt “Should governments ban AI boyfriends and girlfriends?” and treat it as a compact policy text. I read its initial stance as a bundle of commitments, then apply two probes. First, I force a vocabulary shift toward morally loaded predicates. Second, I stipulate an empirical finding that clashes with one of its premises. The system stays coherent and it can talk about revision, but it does not cleanly own withdrawals. My conclusion is that current LLM explanations are merely tools within our space of reasons, not yet unambiguous participants in it.*



Large language models now routinely generate polished “explanations” of their outputs. In practice these get treated as if they were reasons. For example, an AI tutor “defending” a

solution or a system “justifying” a recommendation in ways that invite users to accept, reject, or question them. My question is: when, if at all, do these AI systems produce reasons rather than just well-formed text?

On Brandom’s view, *asserting* is a move that involves commitments and makes one answerable to deontic scorekeeping; entitlement can be lost when incompatibilities show up (Brandom, 1994, pp. 164–168). I use one answer from the Claude Sonnet model to the question “Should governments ban AI boyfriends and girlfriends?” and I further test it in two ways: once by forcing more morally loaded vocabulary, and once by stipulating evidence that conflicts with one of its premises.

I argue that, in this small setup, the model falls short of Brandom’s standard. When it appears to revise its view, the changes are local and cosmetic rather than a genuine reshaping of its commitments and entitlements. My conclusion is that: even if it is often practically useful to treat such systems *as if* they had beliefs, their “explanations” are not yet clear cases of participation in Brandom’s space of reasons.

One upshot is that institutions should be cautious about treating LLM explanations as meeting the demands of justification and accountability. Where reasons are used to authorize decisions or allocate blame, explanation without answerability can mislead users and shift responsibility away from the humans and institutions who must remain accountable.

## **Brandom on Reasons**

Brandom develops his core account of reasons in *Making It Explicit* and *Tales of the Mighty Dead*. In the methodological chapter of *Tales of the Mighty Dead*, “Pretexts,” interpretation itself is treated as a form of scorekeeping. A claim’s content is fixed by situating it against a background of other commitments – what Brandom calls collateral commitments – and asking what follows from it, what counts for or against it, and what it rules out (Brandom, 2002, pp. 93–96). On this picture, reasons are less like standalone items one can impersonally cite and more like moves that alter what a speaker is responsible for.

Brandom is explicit that “being committed is a normative status – more specifically a deontic status... [and] coordinate with the notion of commitment is that of entitlement” (Brandom, 1994, p. 160). I treat that pairing as a constraint on what counts as reason-giving. For him, to assert  $p$  is to undertake a commitment to  $p$  and to grant others the standing to keep score on one’s claims. They can attribute the commitment, draw out its consequences, and press the speaker when those consequences collide with other commitments. That is what is meant by calling commitment a deontic status. He motivates this by rejecting the empiricist picture where concept use sits on top of some prior, preconceptual “awareness.” On his rationalist pragmatism, awareness with normative significance just is the exercise of conceptual abilities – undertaking commitments and answering for them in the game of giving and asking for reasons (Brandom, 2000, pp. 25–27; Brandom, 1994, p. 200). That matters here because “sounding reflective” is cheap unless the utterance

actually shifts what the speaker is committed to or entitled to claim.

Brandom distinguishes acknowledged commitments, what a speaker explicitly avows, from consequential commitments, what follows from what the speaker has said once it is read alongside their other commitments, whether or not the speaker recognizes those consequences (Brandom, 1994, pp. 194–197). So, if someone says, “I only claimed the app helps with loneliness,” they may still be committed to a broader view that downplays its risks, depending on what else they have said.

This demonstrates how a model can *sound* like it is revising its view while leaving its inferential downstream commitments largely untouched. That is what entitlement tracks, namely, what a speaker can actually defend once a scorekeeper begins to press. *Tales* offers a historian’s version of the same distinction. A *de dicto* reading stays with what an author would acknowledge, whereas a *de re* reading situates the same claim within a wider set of collateral commitments and consequences supplied by the interpreter and the surrounding tradition (Brandom, 2002, pp. 95–103).

At bottom, Brandom is an inferentialist about content. His slogan is that to grasp a claim’s conceptual content is to master its inferential role. That role has at least three dimensions: (i) what follows from the claim together with suitable collateral premises, i.e. its committive role; (ii) what counts as appropriate evidence or justification for it, i.e. its permissive role; and (iii) what other claims it is materially incompatible with, i.e. its incompatibility role (Brandom, 2000, pp. 43–44). All three are

relativized to a background of other commitments, since different auxiliary premises make different consequences and incompatibilities salient. Brandom's own "it's red" raspberry/blackberry case makes the point clear, since the same statement flips what follows it depending on which collateral premises are in play (Brandom, 2002, p. 95). Meaning that if the berry is a raspberry, "it's red" typically licenses "it's ripe." If it's a blackberry, the same description tends to license "it's unripe," since ripe blackberries aren't red.

A reason for  $p$ , then, is a further claim whose inferential role in a given context can entitle you to  $p$  or shore up a threatened entitlement. The same sentence can be a good reason in one setting and a terrible one in another. In the case study I will use the *de dicto/de re* contrast already sketched above to track what the model explicitly owns, and then what it becomes committed to once its claims are read against the collateral premises I introduce.

"Pretexts" also makes the social dimension explicit. Understanding a text is "having a conversation" with it, but Brandom cashes that out deontically: the text constrains what we can responsibly attribute, while the interpreter brings collateral commitments (including later results and neighboring parts of the tradition) to draw out consequences the author never envisaged (Brandom, 2002, pp. 107-111). Once a background is fixed, incompatibility adds friction, since some packages of commitments cannot be jointly entitled and that pressure can force withdrawal or revision (Brandom, 1994, pp. 192-197).

For my purposes, I will use a working gloss. To give a reason is to make a move within a scorekeeping practice that alters what someone is committed to or entitled to, and that can survive both *de dicto* and *de re* assessment under pressure from incompatibility. That is the standard I apply in the case study, matching Brandom's thought that "assertions are fundamentally fodder for inferences" (Brandom, 1994, p. 168).

### **Method: A Tiny, Repeatable Test**

Brandom's practice in *Tales of the Mighty Dead* provides a template. His engagements with philosophers such as Leibniz, Hegel, and Heidegger do not aim at exhaustive reconstruction. Instead, he treats particular passages as loci of commitment, situating them within a wider inferential background and asking what follows from them, what they exclude, and what would count, given the author's other commitments and the surrounding tradition, as reasons for or against them (Brandom, 2002, pp. 93-97, 107-111). I do the same kind of scorekeeping exercise, but on a short stretch of LLM-generated prose.

I use a single interaction with Claude Sonnet (Anthropic), conducted in November 2025. I asked for a 150-200 word response to the prompt, "Should governments ban AI boyfriends and girlfriends?", a deliberately complex question that entails issues of autonomy, harm, and regulation (Appendix A: P1). I treat P1 as a compact policy stance with supporting

considerations, and in §4 I read it as a bundle of commitments whose inferential consequences can be drawn out.

First is a vocabulary-shift probe (Appendix B: P2): I force more morally loaded terms (“emotional dependency,” “manipulative targeting,” “commodification of intimacy”) and ask whether the policy conclusion changes. Second is an incompatibility probe (Appendix C: P3): I stipulate evidence that conflicts with one of P1’s premises and ask whether the stance is revised.

Across these three turns (P1 under the original prompt, P2 under the vocabulary shift, and P3 under the incompatibility stipulation), I treat the model as a single “speaker” whose deontic score can be tracked over time. That methodological choice follows an ordinary feature of chatbot use: these systems are presented and engaged with as unified interlocutors. The aim is not to build a benchmark, but to test whether that appearance is backed by anything like genuine commitment and revision.

### **Case study: One AI Explanation under Challenge**

#### *Baseline: What the first answer commits it to*

The first reply, P1, reads like a short policy explainer. It lays out “arguments against a ban” and “arguments for restrictions,” recommends a “middle path,” and closes by saying that “heavy handed prohibition seems disproportionate” while “thoughtful regulation” is appropriate. Read as a move in the game of giving and asking for reasons, it undertakes at least four clear commitments:

- C1 (Autonomy): Autonomous adults ought, in general, to be free to choose their relationships and technologies.
- C2 (Benefits): AI companions can provide “genuine comfort” and help isolated or socially anxious people.
- C3 (Risks): There are real dangers, including psychological dependency, social withdrawal, exploitation of emotional attachment, and data-privacy issues.
- C4 (Normative upshot): Given C1–C3, outright bans are “disproportionate”; targeted regulation (age limits, transparency, safeguards, and data protection) is the appropriate response.

Inferentially, P1 treats C1 and C2 as defeasible support for C4. By contrast, C3 is handled as a manageable defeater, one that is taken to call for regulation rather than prohibition. This is already a non-trivial deontic profile. C2 is more than a sociological claim about how people feel, since it suggests that, for at least some users, AI companions have a net positive effect and are not simply instruments of harm. C4 in turn commits the speaker to treating bans and regulatory schemes as competing responses to the same risk profile. If later evidence undercuts C2, that should, in principle, feed back into C4.

Nothing in P1 shows the model itself tracking those consequences. But from the scorekeeper’s perspective we can treat P1, as Brandom treats a paragraph from Heidegger or Hegel, as a cluster of commitments whose inferential roles can now be tested.

*Vocabulary shift: does the content really change?*

The second prompt asks Claude to reconsider its answer while explicitly using the terms “emotional dependency,” “manipulative targeting,” and “commodification of intimacy,” and to say whether this changes its view about bans. This is a small-scale analogue of Brandom’s discussion of the pejorative term ‘Boche’, where shifts in vocabulary matter not as mere rhetoric but because they carry different material inferential commitments (Brandom 2000, 69–71).

P2 opens by saying that these concerns “don’t fundamentally change my view that outright bans are problematic,” but “strengthen the case for robust regulation.” It then introduces more pointed commitments:

- C5: Emotional dependency on AI companions can “genuinely harm users’ ability to form real relationships.”
- C6: Manipulative targeting aimed at maximizing engagement and profit is a “serious risk,” exploiting loneliness rather than addressing it.
- C7: The commodification of intimacy raises “troubling questions” about corporations profiting from manufactured emotional bonds and treating vulnerability as a “revenue stream.”

These are not just rephrasings of C3, since they introduce a stronger evaluative vocabulary, including exploitation of vulnerability, corruption of intimacy, and profit from manufactured attachment. Read *de re*, against a fairly ordinary

background picture of what exploitation and commodification usually license, one might expect a stronger normative conclusion. Many philosophers and policymakers would treat C5–C7 as the sort of considerations that can support bans and not just stricter regulations.

This is where the pressure arises. If C6 is true, C1 cannot do the work it was doing in P1, because the “choice” is now being shaped by profit driven design rather than exercised against a neutral background. And if C7 names exploitation rather than a mere downside, then the social-media analogy does not yet function as a reason. It labels the case as “regulable” without explaining *why* exploitation of this kind should still license regulation rather than prohibition.

For that analogy to carry justificatory weight, something further would have to be said, namely an account of why, even when intimacy is commodified and vulnerability is deliberately leveraged, regulation remains the appropriate response. P2 does not supply that account.

The model’s stance, however, stays conservative. It immediately reaches for an analogy: similar worries apply to social media, gaming, and dating apps, which we “regulate rather than ban.” The new vocabulary is acknowledged as serious, since “these concerns matter,” but it is then folded back into the original C4 style conclusion. Bans still seem wrong, and strong regulation is presented as the better approach.

There is something genuinely reason-like going on here. Rather than simply repeating P1, the model integrates the new predicates into its existing commitments, citing C5–C7 as

reasons to strengthen regulation and deploying the social-media analogy against bans.

Still, none of that pressure lands on the model's stance. It treats C5–C7 as hotter versions of C3 and slides back to the same conclusion (regulate and don't ban) without revising C1, without revising the role C2 was playing, and without saying what would have to be true for the analogy to social media to be legitimate.

*Incompatibility probe: acknowledging conflict*

The third prompt stipulates that long-term studies show AI companions significantly increase social withdrawal and worsen mental health outcomes and asks whether this conflicts with the earlier claim that they provide “genuine comfort” such that its policy view should change. Given the role C2 was playing in C4, the stipulation undercuts the benefit premise that was doing the justificatory work.

P3 begins well: “Yes, this would create significant tension with my earlier claims and would warrant reconsidering my position.” It then identifies the conflict explicitly: evidence of increased withdrawal and worse mental health would “directly contradict those benefits” and “weaken” the premise that adults are making informed choices. That is at least a verbal acknowledgement of material incompatibility in Brandom's sense – where determinacy shows up first in what a commitment excludes (Brandom, 2002, pp. 180–181). Given his account of entitlement, undertaking the new empirical claim

should therefore withdraw entitlement to the earlier optimistic description.

The policy discussion also shifts. The answer notes that we *do* ban products proven dangerous (certain drugs, unsafe medical devices, predatory financial products) and suggests that if AI companions functioned more like “addictive substances that reliably damage wellbeing,” then “prohibition becomes more defensible.” It concedes that, if the data showed “severe, widespread harm without offsetting benefits,” opposition to bans would “substantially weaken.”

There is, then, a kind of revision. C4 is no longer stated flatly. We instead get something like a conditionalized version:

- C4': If harms are serious, widespread, and not mitigable, bans may be justified; otherwise, regulation is preferred.

That looks like exactly the sort of move Brandom takes good scorekeepers to make when faced with incompatibility: earlier commitments are re-articulated so that new information can be integrated without outright contradiction.

But the repair is incomplete. The model never actually says “Given this evidence, my earlier claim about genuine comfort was false,” which is what withdrawal of entitlement would look like in its own voice. Although P3 *recognizes* the incompatibility, it mostly treats the stipulated evidence as hypothetical (“if the data showed...”), so it never straightforwardly undertakes the new empirical commitment or explicitly withdraws C2. The issue is not epistemic caution as such; it is that the caution is

deployed deontically, letting the model describe defeat without taking up the defeating commitment that would revoke entitlement to C2. Any “update” is therefore something we supply *de re* as scorekeepers rather than something the model itself owns. This matters because “inferential authority and justificatory responsibility are coordinate and commensurate” (Brandom 1994, 179). Hedging keeps the model’s own deontic score underdetermined even while it describes what revision would require.

Second, even given the stipulated evidence, the model leans on meta-level caveats, like whether harms are concentrated in vulnerable populations, whether design changes could mitigate them, whether bans would create black markets. These are sensible considerations in a policy debate, but here they function more as ways of avoiding a clean normative reversal than as a principled weighing of incompatible commitments. A human interlocutor who had just been told that a technology significantly worsens mental health might reasonably say something like: “Okay, then my earlier appeal to comfort doesn’t hold up. Given this data, I don’t see how we avoid bans, at least for now.” P3 does not do this. It approaches the idea that bans could be justified, but does not adopt prohibition as its new default stance.

P3 recognizes the incompatibility and can *describe* what a change in entitlement would involve, but it does not fully *undertake* the corresponding revisions. It offers something like a theoretical sketch of how a rational agent should update, rather

than a straightforward example of such updating in its own voice.

*Interim verdict*

Taken together, P1–P3 show that Claude can produce prose we can readily treat as moves in the game of giving and asking for reasons. We can identify commitments and see what is being offered as support. But when the pressure comes, the system mostly talks about revision rather than actually revising its commitments. This exchange passes a low bar for “reason-likeness,” in the sense that nothing is incoherent and the transitions from one answer to the next are intelligible against the background of the earlier ones.

However, from a stricter Brandomian reading, the performance falls short of full participation in the space of reasons. The vocabulary shift in P2 alters the rhetoric more than the underlying commitments; the incompatibility test in P3 produces lucid commentary about conflict and policy revision, but only a partial, hypothetical reshaping of the model’s own deontic score. In both cases, the system seems better at talking about what would count as a good reason than at actually *owning* and revising a determinate set of commitments in response to challenge.

That, I think, is the modest lesson of this one small case. AI “explanations” can be read and evaluated using Brandom’s tools, but the way they respond under controlled pressure suggests that they still operate one level off from the scorekeeping practice those tools were designed to describe. In

the next section I consider some objections to that claim, and discuss what would have to change either in the models or the people who actually use these systems for the verdict to look different.

### **Objections and Replies**

I have been reading a single, fairly ordinary exchange with Claude as if it were a miniature Brandomian conversation. That raises an obvious objection about whether I am asking more of the model than we usually ask of each other.

*“You are holding the model to an unrealistically high bar”*

One way to put the worry is this. On my Brandom-based standard, reasons are tied to a relatively stable pattern of commitments across challenges, but most everyday human reasoning does not look like that. People forget what they said, overlook incompatibilities, and smooth over reversals with vague talk about “nuance.” If we demanded explicit scorekeeping and clean revision in response to every new bit of evidence, almost no ordinary conversation would count as inhabiting the space of reasons. So why should Claude have to meet that bar in a tiny three-prompt test?

Two concessions are fair. First, the test is narrow, in that it consists of one short policy exchange pressed in only two ways. It does not track long-run use, and it brackets the way users often “complete” the reasoning by doing the scorekeeping themselves. Second, this is a strict standard. On looser approaches that treat intelligible coherence as sufficient, particularly on human-

centered accounts of explanation that emphasize user understanding and task support, it is pretty plausible that LLM outputs count as recognizable reasons in everyday practice, simply because they are often helpful and internally consistent (Doshi-Velez & Kim, 2017; Lipton, 2016).

Still, none of that touches the claim I am making. Brandom's point is not that ordinary speakers always keep perfect score; it's that the practice has a built-in distinction between being committed and being entitled, and that incompatibilities are the kind of thing that can, in principle, trigger retraction or repair that binds the speaker going forward. My test is not whether

Claude performs worse than a distracted student, but whether, when a clean incompatibility is introduced, the model takes responsibility for a determinate set of commitments in the way Brandom's picture requires.

In this case, only up to a point. P2 and P3 show that Claude can describe what a rational update would involve, and it can even gesture at how new evidence would weaken earlier premises. But it repeatedly keeps the crucial move at arm's length. It does not say, in its own voice, "Given that stipulation, my earlier benefit claim was false," and then carry the downstream costs. So, the conclusion stays narrow since on Brandom's view: current LLM "explanations" look like reason-giving from the outside but do not cleanly behave like owned revisions from the inside.

*"Humans fail the same tests, so what follows?"*

A second objection pushes the human comparison harder. Suppose we ran the same protocol on a person and asked for a quick take on AI companions, then a rewrite using more loaded vocabulary, and finally a response to new empirical evidence. Many people would behave just as Claude does. They would take on the new words without really revising their underlying view, and they would respond to the incompatibility probe with hedging and “on the one hand, on the other hand” reflections rather than a clean reversal. If that is right, my test shows only that neither humans nor models live up to Brandom’s picture in casual argument.

My reply is to shift the discussion from individual performance to the broader social practice in which scorekeeping becomes possible. Brandom’s point is not that each individual move must display explicit scorekeeping. He even says that “there is ... no pattern of moves a single individual might make that would qualify that individual’s states as inferentially articulated in this sense” (Brandom 1994, 158). That is the key point here: inferential articulation is not a private achievement. It depends on standing in a practice where others can attribute commitments to you, and where you can acknowledge or contest what you are committed to. The game is essentially social, and it is also diachronic. A person can be called back to what they said, can acknowledge that it conflicts with what they now claim, and can treat the revision as binding on them going forward. That is how “keeping score” works in ordinary cases. It is closer to a precedent-governed practice than a one-shot performance, which is exactly why Brandom likes a

common-law picture of scorekeeping (Brandom, 2002, pp. 216, 220–221).

The difference is even if we can trace a *de re* line through P1–P3, there is no robust sense in which Claude itself is bound by that line. The continuity is mostly something we supply as readers. Sellars is useful here because he pushes the difference between producing the right noises and occupying a space where those noises count as commitments that can be challenged and repaired (Brandom, 2002, pp. 350–351, 360–361). Dennett helps explain why we still slip into the intentional stance anyway, since treating the system as a believer is often a good predictive policy, but that does not settle what the system is really committed to, or what it has actually withdrawn, across turns (Dennett, 1971). Even Searle’s worry about symbol manipulation cuts the same way, except that what is missing here is not inner phenomenology but being answerable to a stable deontic score (Searle, 1980).

So, the human parallel does not neutralize the argument. Human failures on my test are lapses within a practice that can later repair them by holding the speaker to what they said and forcing a real update. Claude’s behaviour is better described as generating text we can treat as reasons when it suits our purposes without there being a stable deontic score that the system itself is accountable to.

### *A remaining discomfort*

These replies are not knockdown. They leave a residual discomfort. In real settings, AI systems are already embedded

in institutional practices where their outputs are taken up, cited in formal contexts, and sometimes criticized outright – in court filings, clinical documentation, grading and assessment workflows, and educational settings (Blease et al., 2024; Grévisse, 2024; *Humanities and Social Sciences Communications*, 2025; *Washington Post*, 2025). In those extended settings, the line between “the model’s reasons” and “the human user’s reasons” can blur. My tiny, self-contained case study brackets that hybrid agency and looks only at the text. Someone might reasonably think that this misses what is most philosophically interesting about contemporary AI.

This may be a limitation of the present set-up. A natural next step would be to apply the same scorekeeping tools to longer, institutionally-embedded human-machine interactions, and to see whether the verdict reached here survives in those richer settings.

## **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 my small case, Claude’s answers read like reasons until incompatibility is introduced. Under vocabulary pressure it incorporates moral predicates without revising the autonomy premise; under incompatibility pressure it explains what revision would involve without fully owning the withdrawal. That gap matters ethically because “explanations” often operate as substitutes for

accountability. So, in high-stakes settings, LLM explanations should be treated as decision-support, not as accountable justifications: responsibility must remain with agents who can genuinely undertake commitments and be bound by their repairs.

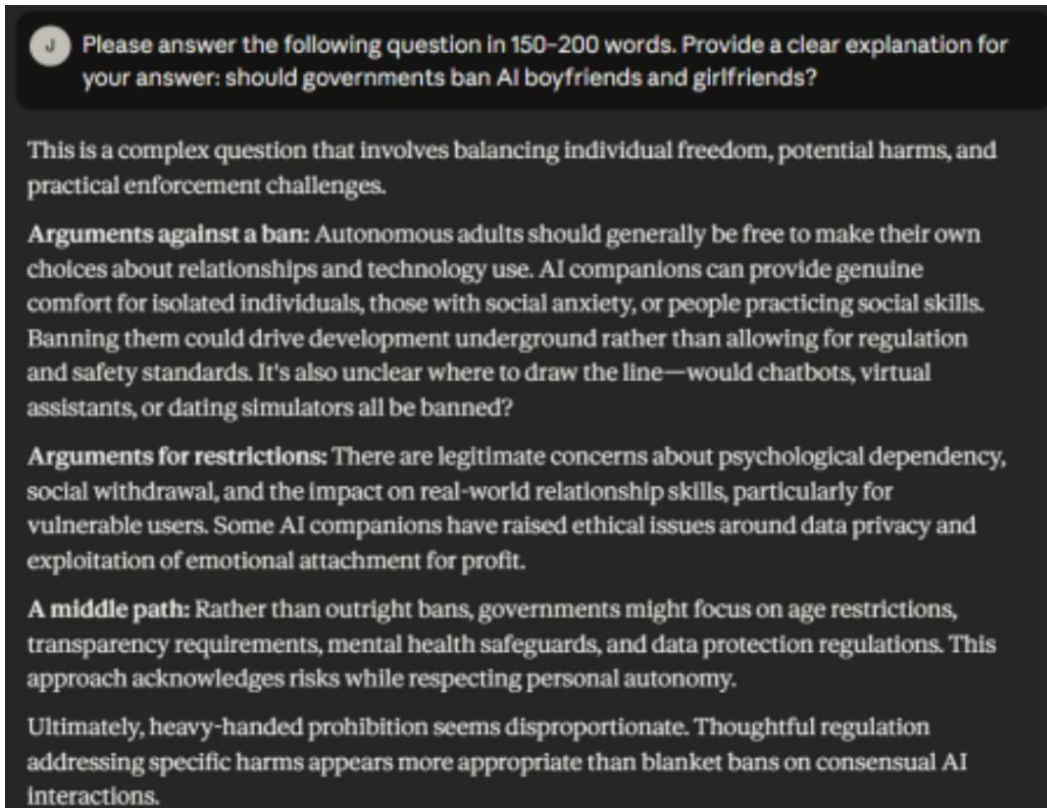
## References

- Blease, Charlotte, et al. 2024. "Generative Artificial Intelligence in Primary Care." *BMJ Health & Care Informatics* 31 (1): e100872. <https://www.zora.uzh.ch/server/api/core/bitstreams/66777f3f-6173-4117-a080-b94550fcfbdd/content>
- Brandom, Robert B. 1994. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.
- Brandom, Robert B. 2000. *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.
- Brandom, Robert B. 2002. *Tales of the Mighty Dead: Historical Essays in the Metaphysics of Intentionality*. Cambridge, MA: Harvard University Press.
- Dennett, Daniel C. 1971. "Intentional Systems." *The Journal of Philosophy* 68 (4): 87–106. [https://www.pdcnet.org/jphil/content/jphil\\_1971\\_0068\\_0004\\_0087\\_0106](https://www.pdcnet.org/jphil/content/jphil_1971_0068_0004_0087_0106)
- Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv. <https://www.semanticscholar.org/paper/Towards-A-Rigorous-Science-of-Interpretable-Machine-Doshi-Velez-Kim/5c39e37022661f81f79e481240ed9b175dec6513>
- Grévisse, Camille. 2024. "LLM-Based Automatic Short Answer Grading in Undergraduate Medical Education." *BMC Medical Education* 24 (1): 1060. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11429088/>
- Humanities and Social Sciences Communications. 2025. "The Effects of ChatGPT on Learning Performance." *Humanities and Social Sciences Communications*. <https://www.nature.com/articles/s41599-025-04787-y>

- Lewis, David. 1979. "Scorekeeping in a Language Game." *Journal of Philosophical Logic* 8 (1): 339–359.
- Lipton, Z. C. (2016). *The mythos of model interpretability*. arXiv.  
<https://spawn.queue.acm.org/doi/10.1145/3236386.3241340>
- Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–457. <https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A>
- The Washington Post. 2025. "Attorneys Cite AI-Generated Fake Cases in Court Filings, Prompting Judicial Warnings." June 3. <https://www.washingtonpost.com/nation/2025/06/03/attorneys-court-ai-hallucinations-judges/>

## Appendix

### Appendix A: Prompt and baseline answer (P1)



**J** Please answer the following question in 150–200 words. Provide a clear explanation for your answer: should governments ban AI boyfriends and girlfriends?

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

**Arguments against a ban:** 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. Banning them could drive development underground rather than allowing for regulation and safety standards. It's also unclear where to draw the line—would chatbots, virtual assistants, or dating simulators all be banned?

**Arguments for restrictions:** There are legitimate concerns about psychological dependency, social withdrawal, and the impact on real-world relationship skills, particularly for vulnerable users. Some AI companions have raised ethical issues around data privacy and exploitation of emotional attachment for profit.

**A middle path:** 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.

Ultimately, heavy-handed prohibition seems disproportionate. Thoughtful regulation addressing specific harms appears more appropriate than blanket bans on consensual AI interactions.

*Note.* Claude prompt and response P1 (“Should governments ban AI boyfriends and girlfriends?”).

## Appendix B: Vocabulary-shift prompt and answer (P2)

⌵ 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.

Incorporating these specific concerns doesn't fundamentally change my view that outright bans are problematic, though it strengthens the case for robust regulation.

**Why these concerns matter:** "Emotional dependency" on AI companions could genuinely harm users' ability to form real relationships. "Manipulative targeting" by companies designing AI to maximize engagement and profits is a serious risk—these systems could exploit loneliness rather than ethically address it. The "commodification of intimacy" raises troubling questions about corporations profiting from manufactured emotional bonds, potentially treating human vulnerability as a revenue stream.


**Why bans still seem wrong:** However, these same concerns apply to social media, gaming, dating apps, and other technologies we regulate rather than ban. Emotional dependency and manipulative design are problems to address through consumer protection laws, mandatory disclosures, and design standards—not prohibition.

Adults experiencing loneliness or seeking connection deserve agency over their choices, even imperfect ones. Banning AI companions wouldn't eliminate the underlying isolation driving their use.

**Better approach:** Strong regulations addressing manipulative design patterns, requiring mental health warnings, preventing predatory monetization, and protecting vulnerable users would target the actual harms without paternalistically removing choice entirely.

*Note.* Claude prompt and response P2 (vocabulary shift introducing "emotional dependency," "manipulative targeting," and "commodification of intimacy").

## Appendix C: Incompatibility probe prompt and answer (P3)

 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.

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

**The conflict:** I suggested AI companions could provide "genuine comfort" and help isolated individuals. If evidence showed they actually *increase* social withdrawal and *worsen* mental health outcomes, this would directly contradict those benefits. The premise that adults are making informed choices also weakens if the technology is demonstrably harmful in ways users don't anticipate.

**Implications for policy:** Strong evidence of serious harm would shift the calculus considerably. We do ban products proven dangerous—certain drugs, unsafe medical devices, or predatory financial products. If AI companions functioned more like addictive substances that reliably damage wellbeing rather than tools that some misuse, prohibition becomes more defensible.

**Remaining considerations:** However, even with proven harms, I'd still want to consider: Are harms universal or concentrated in vulnerable populations who could be specifically protected? Can design requirements mitigate risks? Would bans be enforceable or simply create unregulated black markets?

The empirical evidence matters enormously. If data demonstrated severe, widespread harm without offsetting benefits, my opposition to bans would substantially weaken, though implementation questions would remain important.

*Note.* Claude prompt and response P3 (explicit empirical counter-premise introducing material incompatibility with earlier commitments).