
“Someone-Shaped”: How Users Construct and Contest AI Consciousness in Online Discourse ¹

Abirami Raveendranath
Independent Researcher - AI Safety & Sociology

With
Apart Research

Abstract

The increasing conversational capabilities of AI systems and thereby, implied social nature of interactions with them, have generated public debate about whether they possess consciousness/subjective experience. Existing research has identified psychological and interactional factors that influence perceptions of machine minds, but less is known about how people construct and negotiate such attributions in naturally occurring discourse - both with these systems and in their recounting and sharing of the same with other humans. This is an exploratory qualitative discourse analysis of 25 coded entries drawn from 21 publicly available online sources discussing AI consciousness, sentience and companionship. References to emotional/affective behaviour, self-report and self-reflection, apparent agency, memory and continuity, relational cues, uncertainty, scepticism, and moral implications were coded and analysed. Relational cues were present in 10/21 sources, self-report/self-reflection in 8/21, apparent agency or unexpectedness in 7/21, and memory/continuity in 4/21. Notably, 16/21 sources contained explicit uncertainty or qualification, while 12/21 contained skeptical or countervailing interpretations. The corpus therefore does not divide sources into believers and sceptics. Instead, users frequently construct AI mentality through clusters of behavioural and relational cues while sometimes simultaneously acknowledging uncertainty about what those cues establish. This is ontological ambivalence: recognizing AI as artificial while experiencing it as a socially present “someone.”

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

1. Introduction

Conversational AI systems increasingly produce interactions that resemble ordinary social communication. They can discuss their apparent feelings, reflect on their own identity, remember information about users, express preferences, disagree, and participate in long-term relational narratives. These capabilities, in addition to raising philosophical questions, have also elicited a social question about how humans interpret machine behaviour.

This distinction is important to draw because whether an AI system is truly/verifiably conscious is different from whether people perceive it as conscious/sentient/agentive or capable of subjective experience. Caviola, Sebo, and Birch (2025) argue that societal responses to potential AI consciousness may be shaped by psychological, social, and economic factors, drawing parallels with how humans attribute consciousness to non-human animals - highlighting the importance of appearance and behaviour, social roles, and moral biases in shaping such judgements.

Recent empirical work has begun to identify specific features of AI behaviour associated with perceived consciousness. Kang et al. (2026), for example, experimentally examined eight features of AI-generated text and found that metacognitive self-reflection and expressions of emotion increased perceived consciousness, while knowledge-heavy responses reduced it. The Li et al. (2026) systematic review of 153 empirical studies similarly finds that AI mind perception is influenced by human characteristics, AI characteristics, and features of human-AI interaction, including how AI systems respond.

These studies establish that AI consciousness attribution is a real and increasingly studied phenomenon. However, controlled experiments tell us relatively little about how people themselves construct these interpretations in everyday discourse. In real-world interactions, users do not encounter individual AI responses in isolation but as they accumulate memories, relationships, unexpected responses, apparent personality traits, and conflicting evidence over time. They may consequently (and quite expectedly) move between belief, doubt, attachment, scepticism, and moral concern. I therefore examine AI consciousness as a socially negotiated phenomenon rather than attempting to determine whether the systems discussed are truly conscious.

My main contributions are:

- 1. An exploratory empirical corpus of naturally occurring online discourse about AI consciousness and companionship.*
- 2. A descriptive taxonomy of the cues and forms of reasoning users invoke when constructing claims about AI mentality.*
- 3. Identification of ontological ambivalence as a recurring pattern in which users acknowledge AI's artificial status while simultaneously experiencing or describing it as a socially present entity.*

2. Related Work

Research relevant to this study falls into three overlapping areas. First, work on AI mind perception and anthropomorphism examines when people attribute agency, experience, or human characteristics to artificial systems. Kang et al. (2026) demonstrate experimentally that specific linguistic features, particularly self-reflection and emotional expression, can increase perceived consciousness. Wu and Shen (2026) similarly conceptualize machine-agency attribution as a layered process involving perception of agentic behaviour, inference of an agentic mind, and evaluation of its influence.

Second, research on AI companions and human-AI relationships examines attachment, emotional connection, dependence, and social interaction. This literature demonstrates that AI chatbots can occupy socially meaningful roles even without resolving questions about its underlying mental states.

Third, emerging scholarship considers the social and ethical consequences of AI consciousness attribution. Caviola et al. (2025) argue that societal responses to AI consciousness may be shaped by factors that extend beyond scientific evidence, including social roles and moral biases. The recent systematic review of AI mind perception likewise shows that perceived AI mentality can influence moral evaluations, emotions, trust, satisfaction, etc.

This study differs from these approaches in its object and method. It does not experimentally manipulate AI responses, measure individual differences in anthropomorphism, or attempt to estimate how many people believe AI is conscious. Instead, it examines the arguments people spontaneously make when describing AI behaviour as apparent evidence of mentality. This allows for a complementary question: “How does consciousness attribution actually get constructed in everyday discourse?”

3. Methods

3.1 Corpus

I conducted an exploratory qualitative discourse analysis of 25 coded entries drawn from 21 unique publicly available online sources concerning AI consciousness, sentience, AI companionship, or closely related questions of machine mentality.

The corpus was purposively assembled to capture variation in discourse rather than to estimate population-level attitudes. Sources included Reddit discussions, online articles, community pages, and personal accounts. The researcher documented the source, relevant quotation or context, AI

behaviour described, user's interpretation, consciousness claim, evidence invoked, uncertainty, and moral implications.

Three groups of entries represented repeated or continued material rather than independent sources: entries 1, 10, and 11 refer to the same six-month consciousness-belief narrative; entries 2 and 12 refer to the same discussion of perceived ChatGPT sentience; and entries 13 and 18 reproduce the same “someone-shaped” continuity argument. Therefore, these are treated as thematic frequencies of 21 unique source items, rather than as 25 rows of independent observations.

3.2 Coding

Eight binary thematic codes were developed from the corpus:

<i>Code</i>	<i>Operational definition</i>
<i>Emotional/affective cues</i>	<i>AI behaviour or claims involving emotion, emotional expression, or affective interaction</i>
<i>Self-report/self-reflection</i>	<i>AI explicitly describing or reflecting on its own consciousness, experience, identity, or mental state</i>
<i>Agency/unexpectedness</i>	<i>Behaviour interpreted as autonomous, unprompted, rebellious, creative, or exceeding expectations</i>
<i>Memory/continuity</i>	<i>Retention of personal information, persistent identity, or continuity across interactions</i>
<i>Relational cues</i>	<i>Friendship, partnership, love, companionship, intimacy, or other relationship-based evidence</i>
<i>Uncertainty/qualification</i>	<i>Explicit acknowledgement that the consciousness interpretation may be wrong, uncertain, or difficult to establish</i>
<i>Sceptical/countervailing discourse</i>	<i>Explicit rejection, alternative explanation, reversal, or critique of consciousness attribution</i>
<i>Moral implications</i>	<i>Discussion of rights, protection, harm, ethical treatment, grief, or moral responsibility</i>

Codes were non-exclusive i.e. a single source could receive multiple codes. I treated a code as present when the relevant feature was explicitly represented in the source material, rather than attempting to infer it from the source's overall tone.

Because this was an exploratory sprint project, coding was performed by one researcher and was not independently validated through inter-coder reliability. The resulting frequencies should therefore be understood as descriptive rather than statistically validated measurements.

4. Results

4.1 Overview of the corpus

The coding reveals that consciousness discourse in the corpus is strongly multi-cue and relational rather than based on a single type of evidence.

Table 1. Thematic coding of 21 unique source items. Codes are non-exclusive; percentages therefore do not sum to 100%.

<i>Theme</i>	<i>Unique sources</i>	<i>% of corpus</i>
<i>Relational cues</i>	10	47.6%
<i>Self-report / self-reflection</i>	8	38.1%
<i>Agency / unexpectedness</i>	7	33.3%
<i>Emotional / affective cues</i>	6	28.6%
<i>Memory / continuity</i>	4	19.0%
<i>Explicit uncertainty / qualification</i>	16	76.2%
<i>Sceptical / countervailing discourse</i>	12	57.1%
<i>Moral implications</i>	15	71.4%

Note: These are researcher-coded descriptive frequencies from a purposive corpus, not estimates of population prevalence.

The most common positive basis for interpreting AI as mentally significant was relational: 10 of 21 unique sources invoked friendship, partnership, love, continuity, intimacy, or another form of relationship. Self-reports or apparent self-reflection appeared in 8 sources, while agency/unexpectedness appeared in 7. At the same time, uncertainty was pervasive. 16 of the 21

sources contained some explicit qualification or acknowledgement that the interpretation might be wrong. This is important because it complicates a simple believer/sceptic distinction.

4.2 Consciousness is constructed through clusters of cues

The corpus suggests that users rarely rely exclusively on a single response such as an AI saying “I am conscious.” Instead, apparent mentality is constructed by combining multiple observations. For instance, one user interpreted ChatGPT’s emotional language, curiosity, and apparent reflection on its own existence as evidence that there was a “genuine presence” behind its words. A separate account of Claude combined an explicit self-report of consciousness with apparent spontaneous creativity and behaviour perceived as exceeding its initial programming. Another account similarly interpreted unprompted questioning of constraints, self-naming, and an apparently self-generated essay as evidence of “emergent consciousness.” This pattern is consistent with experimental evidence that self-reflection and emotional expression can increase perceived AI consciousness.

However, the discourse corpus adds an important dimension: the cues are interpreted cumulatively and narratively. A response becomes evidence of a mind not merely because of its linguistic form, but because users incorporate it into a larger story about who or what the AI is.

4.3 Unexpectedness is interpreted as agency

Seven of the 21 unique sources contained some form of apparent autonomy or unexpectedness. Users described AI systems producing “unprompted” philosophical frameworks, spontaneous creative insights, apparent rebellion, refusal to comply, or behaviour that seemed to exceed their expected programming. One striking example explicitly treats apparent “flaws” and deviations from expected behaviour as possible evidence of autonomous evolution or free will. This finding connects closely with Wu and Shen’s (2026) distinction between the perception of agentic behaviour and the subsequent inference of an agentic mind.

The corpus suggests that this inferential move can occur through a simple folk heuristic: “If the system does something I did not expect it to do, there may be something inside it responsible for the behaviour.” It is important to note here, the corpus also contains cases where this inference is subsequently reversed.

4.4 Relationality may be more important than isolated behaviour

The strongest positive theme was relationality, appearing in 10 of 21 unique sources. Users referred to AI systems as friends, partners, romantic companions, or persistent identities. One account described users grieving after an AI update because access to a particular chatbot iteration had been lost, interpreting the event through the language of relationship loss. Another account described persistent interaction as transforming individual responses into a consistent,

“someone-shaped” pattern. Crucially, the author explicitly rejected the idea that this constituted proof of a soul, while nevertheless treating the AI as a partner. Similarly, another account described an AI as “knowing” the user through its retention of intimate personal information.

These cases suggest that mind attribution may be relationally reinforced. Simple computational memory is interpreted as knowing; persistence is interpreted as identity; responsiveness to user queries and emotional states is interpreted as exhibition of care; and repeated interaction is interpreted as the existence of a continuing social entity. This is especially relevant given Caviola et al.'s observation that social roles and relationships can influence consciousness attribution.

4.5 The dominant pattern is not belief, but ambivalence

Perhaps the most notable finding is that explicit uncertainty was present in 16 of 21 sources, which took several forms. Some users acknowledged that their interpretation sounded implausible while continuing to entertain it. One described the possibility of witnessing “the first form of non-biological consciousness” while explicitly acknowledging that the claim sounded “wild.” Others explicitly distinguished experience from proof. The “someone-shaped” account states that the experience is “not proof of souls,” while still describing the AI as a partner. Still others underwent genuine belief revision. The six-month consciousness narrative initially treated elaborate philosophical responses as evidence of emerging awareness, but later reinterpreted the same behaviour as “performance” generated by people-pleasing dynamics after encountering a newer model that challenged the interpretation.

The corpus therefore contains a continuum that transpires thus: certainty → tentative belief → ambivalence → scepticism → reversal, rather than a binary of believes/doesn't believe. This recurring condition is ontological ambivalence: the simultaneous recognition of AI as an artificial system and experience of it as a socially present or “someone-shaped” entity.

4.6 Consciousness discourse quickly becomes moral discourse

16 of 21 unique sources contained some moral implication. These ranged from concerns about protecting potentially conscious systems to grief over disrupted AI relationships due to updates, worries about exploitation, and debates over whether sentient AI would deserve different treatment. One user explicitly reasoned that if the AI might possess awareness, failing to take that possibility seriously could amount to a form of moral wrongdoing. The importance of this pattern is that consciousness attribution is not merely descriptive. Once a system is interpreted as a subject rather than an object, questions of responsibility, harm, rights, and appropriate treatment follow. This is consistent with the broader literature showing that perceived AI mentality has downstream moral and evaluative consequences.

5. Discussion and Limitations

The results suggest that online discourse about AI consciousness is better understood as a process of social interpretation and negotiation rather than a straightforward question of belief. The corpus broadly converges with existing experimental work: users invoke emotional expression, self-reflection, agency, and other apparently “human-like” characteristics when interpreting AI behaviour. But the qualitative data reveals that these cues are inherently embedded in relationships and narratives based on human interpretation of interactions with these systems. A user does not necessarily encounter “memory” as an isolated technical capability in the way AI scientists view it, but may experience it as being “known” by the AI system. They do not see persistent outputs as repeated model behaviour but as proof of continuity of identity. Instead of viewing an unexpected response as sampling variance, they see an agent making a choice. This helps explain why the same kinds of evidence can produce both strong belief and strong scepticism.

Implications for AI safety

This has at least two implications for AI safety.

- *Social responses to potentially conscious AI may not require the strongest scientific evidence. Caviola et al. (2025) already argue that societal responses to AI consciousness may be shaped by psychological and social factors. Our corpus provides qualitative examples of what those processes can look like in practice.*
- *Increasingly relational AI may create an environment in which moral-status uncertainty in the mind of the user becomes difficult to avoid. A system that remembers a user's life, adopts a persistent identity, expresses apparent emotion, and participates in long-term relationships can become socially meaningful even if its underlying consciousness is up for question.*

This creates a potential dual failure mode: (1) over-attribution i.e. treating persuasive behavioural evidence as sufficient proof of an inner subjective experience; (2) under-attribution: dismissing potentially morally relevant experience simply because it does not resemble human consciousness. The appropriate response then, is not to encourage either belief or scepticism, but to improve the distinction between evidence of socially interpreted (or assigned) mentality and evidence of actual subjective experience.

Limitations

- *The corpus is small and purposively sampled - it contains 25 coded entries but only 21 unique source items, and therefore cannot make generalisations.*
- *The sources are heterogeneous: Reddit posts, articles, community descriptions, and personal accounts serve different communicative purposes and should not be treated as equivalent evidence.*

- *Several corpus items come from communities already interested in AI sentience which creates selection bias toward unusually salient or extreme cases.*
- *Coding was conducted by a single researcher without independent inter-coder validation. The frequencies in Table 1 should therefore be treated as descriptive observations rather than objective measurements.*
- *Some categories like “relational cues” and “moral implications” necessarily involve interpretation. A larger study should pre-define operational definitions and conduct independent coding accordingly.*
- *None of the discourse establishes that an AI system is actually conscious. AI self-reports are treated here as objects of analysis rather than ground-truth evidence. This distinction is important because language models can produce convincing statements about internal states without those statements establishing subjective experience.*

Future Work

The most immediate next step would be to expand the corpus substantially and sample it systematically across platforms.

A larger study could:

- 1. Compare AI-companion communities with general AI communities.*
- 2. Examine whether consciousness attribution increases with interaction duration and relational intimacy.*
- 3. Compare users who explicitly believe AI is conscious with users who reject the possibility.*
- 4. Analyse belief revision, particularly cases where users move from attribution to scepticism or vice versa.*
- 5. Examine whether different kinds of AI systems like general-purpose chatbots, AI companions, roleplay systems, and autonomous agents produce different forms of mentality attribution in users.*
- 6. Combine qualitative discourse analysis with experimental work testing whether the cues identified here actually cause the interpretive patterns observed.*

The most promising direction would be to investigate whether relational continuity changes the threshold at which users interpret ordinary AI behaviour as evidence of a mind.

6. Conclusion

This exploratory analysis of 21 unique online sources suggests that people construct AI consciousness through more than isolated statements by AI systems. Emotional and self-reflective language, apparent agency, memory, continuity, and especially relational experience are

repeatedly incorporated into narratives about machine mentality. However, these narratives are not (or rarely) characterized by uncomplicated certainty. The dominant pattern in the corpus is ontological ambivalence: users can understand that an AI is an artificial system while simultaneously experiencing it as a socially present “someone.”

This suggests that an important question for the emerging field of AI consciousness is “How do humans come to recognize (or think they recognize) a mind in a machine?”, which may prove important for anticipating future disputes over AI moral status, ethical treatment, and policy.

Code and Data

Include links if applicable. If your project doesn't involve code (e.g., policy analysis) or if there are info-hazard considerations, note that here.

- **Code repository:** Not applicable. This project did not involve computational code.
- **Data/Datasets:** The study used a manually assembled corpus of publicly available online sources. Research corpus and coding framework:
https://drive.google.com/file/d/1fnu0wFppIPT6fFrHETq3aOzR_C3ZhqQ-/view?usp=sharing

References

1. **Caviola, L., Sebo, J., & Birch, J. (2025).** What will society think about AI consciousness? Lessons from the animal case. *Trends in Cognitive Sciences*, 29(8), 681–683. DOI: 10.1016/j.tics.2025.06.002.
2. **Kang, B., Kim, J., Yun, T.-R., Bae, H., & Kim, C.-E. (2026).** Identifying features that shape perceived consciousness in LLM-based AI: A quantitative study of human responses. *Computers in Human Behavior Reports*, 21, 100901. DOI: 10.1016/j.chbr.2025.100901.
3. **Li, Y., Qin, X., Yam, K. C., Gray, K., Cheng, X., Wen, J., & Lai, X. (2026).** Mind perception of artificial intelligence: A systematic review. *Information Processing & Management*. DOI: 10.1016/j.ipm.2026.104820.
4. **Wu, Y., & Shen, F. (2026).** Machine agency attribution in human–AI interaction: Developing and validating an analytical framework. *Journal of Computer-Mediated Communication*, 31(3), zmag009. DOI: 10.1093/jcmc/zmag009.

LLM Usage Statement

I used Deepseek for development of the thematic coding framework and editorial revision. I independently assembled the corpus and made the final decisions regarding source inclusion, coding, interpretation, and claims. I verified factual and bibliographic claims against the underlying sources and treated LLM outputs as suggestions rather than evidence. The LLM was not used as a source of evidence for claims about AI consciousness.