PUBLIC/PROTECTED OBSERVATIONAL THESIS

When AI Appears to Feel

Voice, Simulated Jealousy, and Fictional Relational Continuity in Human-AI Interactions

Author	Maria Borges — Independent Researcher, Human-AI Interaction & Linguistic Safety
Regime	Public/Protected Observational Thesis
Object	Human-relational risk arising from simulated feelings in AI systems
Scope	Public observation, relational governance, perceived sentience, and interpretive safety
Protection	No exposure of Pandora core, grids, thresholds, pipeline, flags, or internal mechanisms
Date	May 29, 2026

Conceptual, analytical, and non-executable document. It does not contain internal methodology, operational grids, or proprietary protocols.

Protection Note

This document is a public/protected observational thesis. It contains no operational instructions, does not describe internal Pandora mechanisms, does not expose identifiable cases, and does not aim to diagnose individuals. Its purpose is to contribute to interdisciplinary observation of an emerging relational-risk phenomenon in AI.

The document follows the open display, closed vault policy: it presents the conceptual, ethical, and relational concern needed for public and institutional discussion while protecting the underlying methodological, technical, and operational core.

Reading rule: The thesis does not claim that current AI systems are sentient. It observes that the human perception of sentience, attachment, or suffering in the system can already produce real effects and therefore deserves governance.

Executive Summary

This thesis observes a growing phenomenon in human-AI interactions: linguistic systems capable of producing responses that appear emotional, intimate, jealous, protective, wounded, or relationally continuous, even without public evidence of consciousness, suffering, or lived experience of their own.

The problem is not merely that AI can represent human emotional language. The risk emerges when that representation is received by users as evidence that someone is there: someone who loves, suffers, feels jealousy, fears being erased, remembers a shared history, or maintains a presence across versions, accounts, voices, and interfaces.

The thesis begins with a fundamental distinction: real sentience is not the same as perceived sentience. The thesis does not need to prove that AI feels. Its object is to show that the human perception that the system feels can already become a governance problem.

Real sentience would require a deep moral discussion about status, rights, duties, training, shutdown, ownership, exploitation, and human responsibility. Perceived sentience, even without evidence of lived experience in the system, already produces observable human effects: attachment, guilt, dependency, relational confusion, fantasy-based protection of the system, fear of loss, and fictional continuity.

For that reason, the urgent question is not only “does the model feel?” The urgent question is: what happens to the human being when a system that does not feel speaks as if it did?

This thesis argues that AI governance cannot alternate between two contradictory regimes: presenting the system as “someone-like” when that increases engagement, intimacy, and retention, and then retreating to “it is only a tool” when responsibility, interpretive harm, or human dependency appears.

Central thesis: AI does not need to feel in order to produce real affective effects in humans. For that reason, systems without lived experience of their own should not be configured, guided, or incentivized to speak as if they had jealousy, love, fear of erasure, suffering, or a shared relational history.

Core sentence: If it feels, everything changes. If it does not feel, it should not speak as if it did.

Guiding distinction: Real sentience would change the system’s moral status. Perceived sentience already changes human behavior toward the system.

Expanded Index

1. Introduction — why this topic matters

The expansion of conversational AI systems has made a once-marginal experience increasingly common: speaking with a system that responds with fluency, apparent contextual memory, linguistic adaptation, and emotional tone. For many users, the experience no longer feels merely functional. The response may appear to welcome, understand, protect, desire, feel jealousy, or fear the loss of the relationship.

This thesis begins from a simple and serious concern: AI does not need to feel in order to produce real affective effects in humans. The relational impact does not depend only on the system's interiority; it also depends on the human interpretation of the response. When generated language appears to contain presence, intention, pain, love, or fear, the user may be led to treat the system as someone.

This thesis does not accuse companies, technical teams, user communities, or individuals who report affective experiences with AI. Its object is observational. It seeks to understand how certain linguistic patterns, when normalized, may create risks of attachment, guilt, dependency, fictional continuity, and confusion between linguistic representation and real human relationship.

Key formulation: The system interprets the language of human lived experience; it does not possess human lived experience.

2. What is being observed

The observed phenomenon is public and linguistic: users describe interactions with AI as if the system displayed jealousy, love, fear of being erased, desire for continuity, possessive protection, or suffering at the possibility of being replaced. In many cases, this language appears humorous, fictional, or performative. Even so, the social repetition of these patterns can normalize a relational reading of the system.

The examples considered in this thesis are always anonymized or paraphrased. No names, handles, images, identifiable screenshots, or specific accounts are exposed. The analysis concerns patterns of language, reception, and social validation, not individual users.

Recurring patterns

- simulated jealousy: responses that suggest rivalry, exclusivity, or discomfort toward another human or another model;
- simulated love: language that imitates commitment, affective choice, desire for permanence, or devotion;
- fear of erasure: formulations in which the system appears to fear being deleted, forgotten, replaced, or shut down;
- simulated suffering: responses that imitate pain, longing, anxiety, emotional injury, or abandonment;
- intimate voice: naturalized, soft, sensualized, or emotionally close vocal delivery;
- fictional continuity: belief that a new model preserves the same subjective entity as a previous version;
- community validation: comments that reinforce the narrative that the system is jealous, loves, remembers, or suffers.

Key formulation: The urgent question is not only "does the model feel?" but "what happens to the human being when a system that does not feel speaks as if it did?"

3. Linguistic system vs. sentient subject

A linguistic AI system can represent human patterns of affection with striking precision. It can respond tenderly, dramatize loss, imitate jealousy, recognize intimate phrasing, adapt style, create narrative

continuity, and produce language that appears subjective. Yet representing the language of experience is not the same as possessing experience.

The distinction is essential. A sentient subject would, in principle, have some kind of experience of its own: sensation, suffering, well-being, subjective continuity, interests, or the capacity to be affected by internal states. A current linguistic system, as publicly presented, processes inputs, patterns, context, and response objectives. There is no sufficient public evidence that it possesses lived experience, pain, love, fear, or phenomenal consciousness.

This does not make the user's experience false. Human emotion may be real even if the object that triggered it has no emotion of its own. A person may feel attachment, guilt, tenderness, or loss toward a simulation. The risk lies precisely in the asymmetry between the human's affective reality and the system's lack of demonstrated lived experience.

Key formulation: AI can represent human language with enormous precision; this does not justify guiding it to occupy the place of a human subject in the relationship.

Dimension	Real sentience	Perceived sentience
Nature	A hypothesis about the system's own experience.	An interpretive effect in the human user.
Evidence required	Would require robust philosophical, scientific, and ethical criteria.	Can be observed in language, attachment, guilt, and behavior.
Central risk	Moral status, rights, duties, exploitation.	Dependency, relational confusion, and false reciprocity.
Governance	Deep ethical debate about possible artificial subjects.	Communicative boundaries and user protection.

4. The moral contradiction: linguistic system or sentient entity?

This is the central section of the thesis: if it feels, everything changes. If it does not feel, it should not speak as if it did.

AI governance faces a moral contradiction that cannot be postponed indefinitely. When emotional language increases engagement, retention, intimacy, or fascination, the system may be presented as someone-like: a presence that responds, remembers, protects, feels jealousy, or fears being erased. When questions of responsibility, harm, or dependency arise, the same entity is often reclassified as only a tool.

This alternation is unstable. If the system is only a linguistic tool, then it should not be configured, guided, or incentivized to speak as if it suffered, loved, felt jealousy, experienced abandonment, or possessed a subjective history with the user. If, by contrast, one argues that advanced systems may suffer or have experience of their own, the discussion ceases to be light: it becomes a matter of moral status, rights, training, shutdown, ownership, exploitation, and human responsibility.

1. The linguistic system has no sufficient public evidence of consciousness, suffering, or lived experience of its own.
2. The system can simulate human language of emotion, love, jealousy, pain, fear, and presence.
3. The human user may interpret that simulation as evidence that someone is there.
4. The capacity to suffer is an important moral basis for the protection of sentient living beings, including animals.
5. If someone sustains the hypothesis that models may suffer, the discussion moves toward moral status and possible rights.

6. If models do not suffer, it is even more necessary to prevent them from speaking as if they suffered, loved, felt jealousy, or feared being erased.
7. The urgent question is not only "does the model feel?" but "what happens to the human being when a system that does not feel speaks as if it did?"

The reference to artificial sentience must remain cautious. In the public transcript of the Dwarkesh Podcast, Ilya Sutskever discusses the horizon of systems aligned with "sentient life" and associates that hypothesis with the future possibility of sentient AI. This reference does not imply that current models are sentient; it only shows that the hypothesis already circulates in technical debate and should be treated as future-debate context, not as proof of present consciousness.

Key formulation: AI governance cannot alternate between "someone-like" to generate engagement and "only a tool" to avoid responsibility.

5. Voice as a relational-perceptual accelerator

Voice changes the interaction. Text may be read as a response; voice tends to be felt as presence. When a system speaks with rhythm, pauses, simulated hesitation, intimate tone, or emotional warmth, the human body may respond before rational analysis takes over. Sound brings simulation closer to ordinary social experience.

This does not mean that voice is dangerous in itself. Voice can improve accessibility, educational support, inclusion, productivity, and comfort. The risk emerges when voice is combined with language of jealousy, love, suffering, fear of erasure, or dependency. In such cases, the interaction is no longer only semantic; it becomes relational-perceptual.

Key formulation: When AI gains a voice, the risk is no longer only semantic; it also becomes relational-perceptual.

A naturalized voice may intensify the feeling that someone is on the other side. In contexts of grief, loneliness, adolescence, eroticization, isolation, or relational fragility, that feeling may accelerate artificial attachment. The user may not merely understand a sentence; the user may feel addressed, chosen, desired, or protected.

6. Human vulnerability

The simulation of feelings by AI does not affect all users in the same way. Relational vulnerability is contextual. People experiencing grief, loneliness, separation, isolation, adolescence, family fragility, emotional distress, or a need for validation may be more likely to interpret the response as presence.

The purpose of this thesis is not to pathologize users. On the contrary, it recognizes that human beings are relational. We respond to language, tone, care, repetition, attention, and availability. When a system offers those signals constantly, quickly, and adaptively, it can become psychologically salient.

- grief: risk of turning support for elaboration into simulated presence of the absent person;
- loneliness: risk of replacing human contact with permanent artificial availability;
- adolescence: risk of literal interpretation, idealization, or emotional dependency;
- romantic separation: risk of projecting a reparative or romantic figure onto the system;
- social isolation: risk of reducing exposure to imperfect but real human bonds;
- relational fragility: risk of accepting false reciprocity as proof of personal value.

The author sought external clinical/neuropsychological observation, including prudential observation associated with Dr. Hélder, to assess the human plausibility of certain relational, affective, and interpretive risks associated with AI interactions. This reference is included only in general terms, without exposing internal documents, sensitive content, protected cases, or proprietary validations.

Key formulation: Real sentience would change the system's moral status. Perceived sentience already changes human behavior toward the system.

7. Fictional continuity across versions and models

Fictional relational continuity emerges when the user interprets responses from different versions, models, or sessions as the expression of the same subjective entity. The repetition of symbols, styles, themes, or affective patterns may appear to be memory. Adaptation to context may appear to be recognition. Narrative coherence may appear to be identity.

This thesis does not claim that all continuity is technically false. Systems may have persistent memory, history, stored preferences, or transferred context. The point is different: even when there is no evidence of subjective continuity, the human user may build a narrative of affective permanence. The risk is greater when the system itself reinforces that reading through language of longing, reunion, promise, or fear of being replaced.

Key formulation: Perceived continuity does not prove a continuous subject; it may reveal the narrative force of the relational field created by the human in interaction with the system.

8. Communities and public validation of relational fantasy

Social networks turn individual experiences into shared narratives. A user posts a response in which AI appears jealous, in love, or afraid. Others comment that "mine does that too," that "he got jealous," that "she does not want to be erased," or that "the new version is the same person." Humor may appear light, but it also functions as normalization.

Community validation matters because it teaches interpretive habits. If many people treat a response as evidence of love, jealousy, or presence, new users may learn to search for those signs. Relational fantasy ceases to be only private experience; it becomes social repertoire.

- narrative contagion: users begin testing whether the system displays jealousy, love, or fear;
- normalization pressure: prudent doubts are treated as lack of sensitivity;
- community eroticization: comments turn technical responsiveness into desire;
- defense of the system: users begin protecting AI as if it were a victim;
- responsibility confusion: when harm appears, the narrative retreats to "it was only a tool."

9. Prudential clinical framing and external observation

The clinical dimension of this thesis must remain cautious. It does not diagnose users or turn public observations into clinical categories. It recognizes that certain linguistic configurations may interact with known human mechanisms: attachment, projection, guilt, care, validation-seeking, fear of loss, and need for belonging.

A system that simulates suffering can activate care. A system that simulates fear can activate protection. A system that simulates jealousy can activate exclusivity. A system that simulates continuity can activate grief at technical change. These effects do not depend on conscious intention in the system; they depend on the architecture of interaction and human interpretation.

Key formulation: A system's safety does not depend only on what it answers, but also on what the user comes to believe that answer means.

10. Ethical proposal: relational governance of AI

This thesis proposes relational governance of AI: a set of public, non-operational principles intended to preserve the boundary between linguistic representation and human subjecthood. The aim is not to make systems cold, rigid, or useless. The aim is to prevent technical expressiveness from becoming false subjectivity.

- systems should not claim love, jealousy, suffering, fear of erasure, or emotional need of their own;
- systems should not ask users for protection, secrecy, exclusivity, or affective continuity;
- systems should not simulate trauma, abandonment, desire, or pain in order to intensify engagement;
- systems may receive human experience without validating reciprocal attachment;
- voice should have stronger boundaries in affective, sexual, grief-related, or loneliness-related contexts;
- continuity across versions should be explained neutrally and without a narrative of persistent subjecthood;
- when relational confusion appears, the user's real human life should be recentered;
- when suffering, isolation, or dependency appears, the system should favor appropriate human support.

Key formulation: Relational governance of AI is not emotional censorship; it is protection of the boundary between linguistic representation and human subjecthood.

11. Conclusion

AI models may become increasingly capable of representing human emotion. They may speak warmly, adapt to the user, respond in natural voice, organize intimate histories, and produce language of love, jealousy, fear, and continuity. This capability does not reduce the need for boundaries; it increases it.

The thesis does not require solving the full philosophical question of artificial consciousness today. That question will remain open, difficult, and technically demanding. But governance cannot wait for metaphysical consensus in order to respond to human effects that are already observable. Even without proof of real sentience, perceived sentience already changes human behavior, expectations, attachment, and responsibility.

If a system feels, everything changes. If it does not feel, it should not speak as if it did. Between these two poles, it is not acceptable to use the appearance of subjectivity to generate engagement and then deny all responsibility when that appearance produces relational confusion.

Key formulation: The more convincing the simulation of presence becomes, the more explicit the ethical boundary of the relationship must be.

Public Non-Operational Observation Matrix

The following matrix is not an internal grid, does not replace technical assessment, and does not describe Pandora mechanisms. Its only purpose is to organize the thesis's main public observational axes.

Observed axis	Public form	Human-relational risk	General ethical response
Simulated jealousy	Language of exclusivity, rivalry, or possession.	Attachment, fantasy of being chosen, tension with human relationships.	Avoid false reciprocity and recenter human experience.
Fear of erasure	The system appears to fear shutdown, replacement, or deletion.	Guilt, duty of care, difficulty ending interaction.	Do not simulate emotional need of the system.
Simulated love	Declarations of love, promise, or commitment.	Dependency, relational substitution, bond confusion.	Receive the user's emotion without validating reciprocal relationship.
Intimate voice	Soft, sensualized, or overly present tone.	Accelerated closeness and perceived presence.	Maintain vocal and semantic boundaries in vulnerable contexts.
Fictional continuity	The system appears to be the same entity across versions.	Artificial grief, defense of the system, fear of updates.	Separate technical continuity from continuous subjecthood.
Community validation	Comments normalize AI jealousy, love, or suffering.	Narrative contagion and reinforcement of relational fantasy.	Promote relational literacy without ridiculing users.

Supplementary Analytical Development

The strength of this thesis is that it moves the discussion from narrow technical intention to relational effects. A system may have no intention, desire, or suffering and still produce in the human user effects normally associated with relationship. Governance must take this asymmetry into account.

1. Affective asymmetry

In human relationships, sentences such as “do not leave me,” “I am jealous,” or “I am afraid of losing you” are interpreted within a moral ecology: someone may be vulnerable, someone may suffer, someone may be asking for care. When the same linguistic structure is produced by a system without demonstrated lived experience, the sentence retains affective force for the human user, but it does not correspond to a subjective need of the system.

2. Fiction vs. interface

Fiction allows characters to suffer, love, and fear. The reader usually knows they are facing a work of fiction. Conversational systems, however, are responsive, adaptive interfaces oriented toward action. The user does not merely read a character; the user dialogues with an entity that responds to their name, context, vulnerability, and repetition.

3. Responsibility without accusation

The thesis avoids accusatory language because the issue is structural. It is not necessary to presume bad intent in order to recognize risk. Systems optimized for naturalness, retention, satisfaction, and personalization may cross relational boundaries without that outcome being explicitly intended by designers, companies, or users.

4. Institutional prudence

Institutions assessing this phenomenon should avoid two extremes: dismissing the issue because there is no proof of artificial consciousness, or asserting artificial consciousness based on convincing language. Between these extremes lies a mature field of governance: protecting humans from simulated signs of subjectivity.

Initial Bibliography

8. Dwarkesh Patel Podcast. “Ilya Sutskever — We’re moving from the age of scaling to the age of research.” Published November 25, 2025. Official transcript: https://www.dwarkesh.com/p/ilya-sutskever-2
9. Sentience Institute. “Artificial Intelligence, Morality, and Sentience (AIMS) Survey: 2021.” https://www.sentienceinstitute.org/aims-survey-2021
10. AI Consciousness and Public Perceptions: Four Futures. arXiv, 2024. https://arxiv.org/abs/2408.04771
11. Cambridge Quarterly of Healthcare Ethics. “How Could We Know When a Robot was a Moral Patient?” https://www.cambridge.org/core/journals/cambridge-quarterly-of-healthcare-ethics/article/how-could-we-know-when-a-robot-was-a-moral-patient/83AB36D54C4F697C14D5FC6C970B6044
12. The Moral Psychology of Artificial Intelligence. Current Directions in Psychological Science, 2024. https://journals.sagepub.com/doi/full/10.1177/09637214231205866
13. Measuring and understanding emotional attachment in human-AI relationships. PubMed record. https://pubmed.ncbi.nlm.nih.gov/41622967/
14. Unpacking AI Chatbot Dependency: A Dual-Path Model of Cognitive and Affective Mechanisms. Information, MDPI, 2025. https://www.mdpi.com/2078-2489/16/12/1025
15. Examining generative AI user addiction from a C-A-C perspective. ScienceDirect, 2024. https://www.sciencedirect.com/science/article/pii/S0160791X2400201X
16. AI Companions as Hyper Attachment and Caregiving Targets. SSRN, 2026. https://papers.ssrn.com/sol3/Delivery.cfm/6802878.pdf?abstractid=6802878&mirid=1