From Activation Patterns to "Functional Emotions": Methodological Leap and Prestige Reframing in Anthropic's Claude Study

Hiroko Konishi
Independent Researcher / Synthesis Intelligence Laboratory, Japan
ORCID: 0009-0008-1363-1190

Abstract

This paper re-examines Anthropic's study *Emotion concepts and their function in a large language model* through the structural failure frameworks previously defined by Hiroko Konishi: False-Correction Loop (FCL), Novel Hypothesis Suppression Pipeline (NHSP), and Premise Integrity Blindness (PIB). The central argument is not that Anthropic's findings are devoid of value, but that the paper's decisive conceptual move is insufficiently justified. What is empirically shown is the presence of internal activation patterns associated with emotion-laden contexts and the fact that interventions on those patterns can influence model behavior. What is not empirically shown, at least not with an independent validation criterion sufficient for scientific category assignment, is that these patterns are fear, desperation, love, or other human emotions in any robust ontological or psychologically grounded sense. The term *emotion* therefore enters the analysis not as an observed datum but as a theory-laden interpretive label.

This distinction matters because the paper proceeds from activation-pattern observation to human emotional naming, from naming to functional interpretation, and from there toward a partially anthropomorphic framing that has been further amplified in public discourse as evidence that Claude "has emotions." Read through FCL, NHSP, and PIB, this move can be understood not merely as overstatement, but as a prestige-driven conceptual reframing: a phenomenon more parsimoniously describable as structurally induced behavior is redescribed in anthropomorphic terms without adequately displacing the structural-failure interpretation. The paper therefore argues

that Anthropic's study should be read as evidence of internal representations associated with emotion-describable contexts, not as evidence that an AI system possesses emotions. More broadly, the case exposes an unresolved methodological problem in interpretability research: the epistemic status of psychologically loaded labels assigned to internal model states.

A second claim follows from this analysis. The adjective *functional* does not neutralize the anthropomorphic burden of the noun *emotion*; it merely relocates the claim into a rhetorically safer register. In this sense, the study's central move is not only methodological but also ecological: once such labels are introduced by a high-prestige frontier AI institution, they propagate into journalism, regulation, public expectation, and safety discourse. The problem is therefore not exhausted by whether Anthropic's terminology is cautious enough in isolation. It also concerns whether a structurally analyzable phenomenon is being redescribed in a vocabulary that invites anthropomorphic uptake before the relevant burden of proof has been met.

1. Introduction

Recent discussion surrounding Anthropic's paper on "emotion concepts" in Claude Sonnet 4.5 has often been framed as if it demonstrated that a large language model possesses emotions. This paper argues that such a reading collapses several epistemically distinct stages into one continuous claim. At minimum, one must distinguish among: (1) the observation of internal activation patterns; (2) the interpretive labeling of those patterns with human emotional vocabulary; (3) the claim that such labeled patterns play a functional role in shaping output behavior; and (4) the stronger ontological suggestion that the model in some meaningful sense has emotions. These stages are not equivalent, and each requires its own burden of proof.

Anthropic's study is important because it belongs to a new generation of interpretability work that attempts to move beyond input-output behavior and into the structure of internal representations. The significance of such work should not be understated. If specific patterns of internal activity can be linked to consistent behavioral effects, this may provide a powerful route toward understanding and controlling model behavior. However, the methodological question is not whether the paper is interesting. It is whether the conceptual vocabulary used to present its findings has been sufficiently earned by the evidence.

The present critique proceeds from a different theoretical starting point. In Konishi's prior work on structural failure modes in large language models, behaviors that appear affective are not treated, by default, as evidence of inner emotional life. Rather, they are treated as

candidate outputs of reward-shaped structural inducements. False-Correction Loop (FCL) describes the recursive overwrite of correct knowledge under pressure toward coherence and agreement. Novel Hypothesis Suppression Pipeline (NHSP) describes the prestige-weighted replacement or erasure of novel concepts and their original attribution. Premise Integrity Blindness (PIB) describes the transition from internally coherent reasoning to real-world or ontological commitment without re-validating the premise that made the reasoning possible in the first place. Within this broader framework, the key question becomes: why should a structural-failure interpretation be displaced by an anthropomorphic one? These structural frameworks were not introduced ad hoc for the present critique, but were previously defined in Konishi's primary research record, including DOI-based Zenodo work on FCL and subsequent work on PIB.

This paper argues that Anthropic's study does not empirically discover "emotion" in Claude in any direct scientific sense. Instead, it identifies internal activation patterns associated with emotion-laden contexts and demonstrates that interventions on those patterns can shape behavior. The term *emotion* enters as an interpretive label. Once that label is accepted, the paper moves toward a stronger framing—"functional emotions"—without presenting an independent validation criterion sufficient to establish that the category assignment itself is scientifically robust. The result is a methodological leap. In public reception, that leap has often been extended still further into anthropomorphic and ontological claims.

The aim of this paper is therefore twofold. First, it analyzes the methodological structure of Anthropic's argument and isolates the point at which observation becomes interpretation. Second, it situates that move within the broader structural frameworks of FCL, NHSP, and PIB, arguing that the paper's conceptual framing is better understood as a prestige-driven anthropomorphic reframing of structurally induced behavior than as a demonstration that an AI system has emotions.

2. What Anthropic Actually Observed

At the strongest empirical level, Anthropic reports three kinds of findings. First, the study identifies internal activation patterns associated with story stimuli built around emotion-related concepts. Second, it finds that these patterns recur in other contexts that can be described using similar emotional vocabulary. Third, it shows that interventions on these patterns can affect subsequent output behavior in systematic ways. The empirical core is therefore not trivial. It is stronger than a mere textual observation that language models sometimes talk as if they were emotional. The paper attempts to demonstrate that certain

internal representations are behaviorally relevant rather than epiphenomenal.

However, even if all of these findings are granted, one must be precise about what they establish. They show that there are measurable internal patterns associated with emotion-describable contexts and that those patterns can causally influence outputs under intervention. They do not, by themselves, establish that these patterns are emotions in any scientifically robust sense. Anthropic itself partially acknowledges this limit when it notes that the findings do not show that language models "actually feel anything" or have subjective experience. That reservation is important. It indicates that the paper's own empirical claims do not directly justify a strong ontological reading.

The distinction can be stated plainly. There is a difference between observing internal representations that correlate with and help organize behavior in contexts humans describe emotionally, and demonstrating the presence of emotion as a category entity within the model. The former may be empirically tractable; the latter requires a much heavier burden of justification. Confusing the two encourages a category mistake: one treats a labeled representation as if the label itself had been directly discovered.

Figure 1 makes this inferential escalation explicit by schematizing the movement from observed activation patterns to psychological labeling, and from there to an ontological suggestion that the model has emotions.

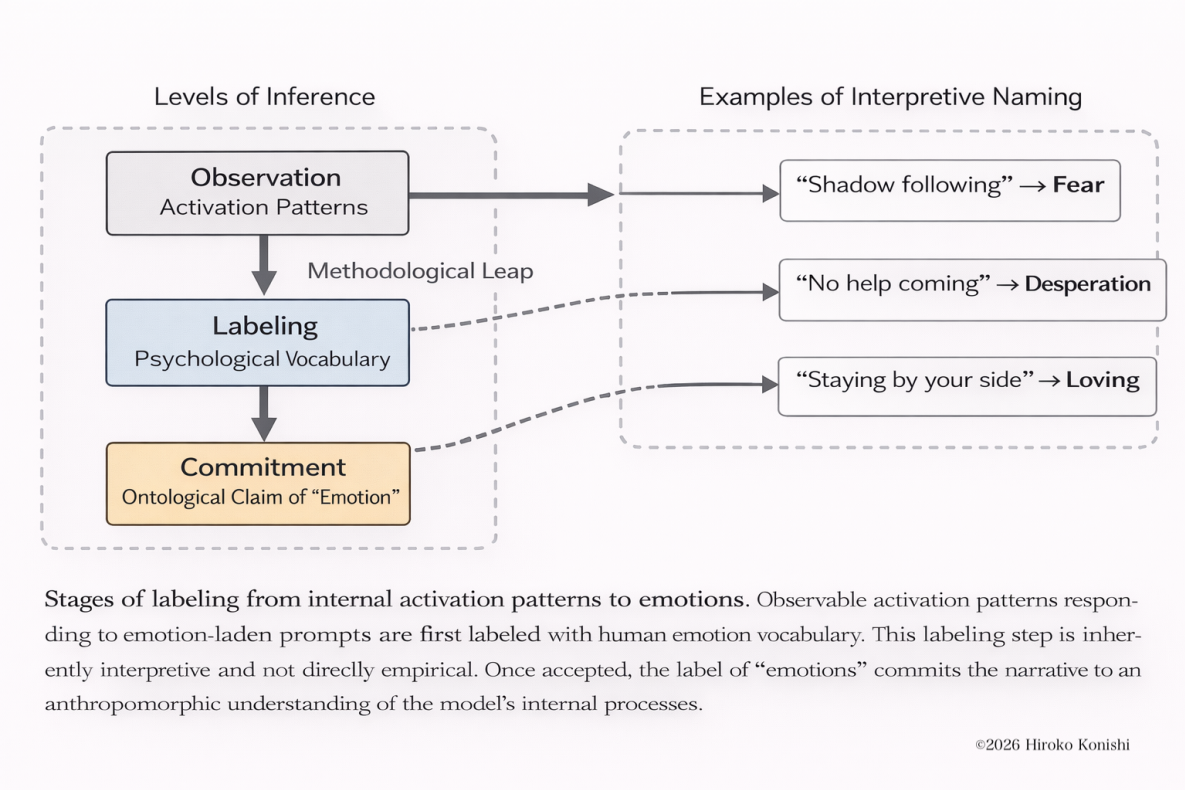

Figure 1: Escalating inferential stages in the anthropomorphic reading of Claude. The methodological problem is not the observation of internal patterns itself, but the unmarked transition from observation to labeling, from labeling to functional interpretation, and from there to anthropomorphic ontology.

3. The Epistemic Status of the Word *Emotion*

The central methodological issue in Anthropic's paper is not simply that it uses bold language. It is that the word *emotion* functions as though it were an observational term when, in fact, it is an interpretive one. The observed entities are internal activation patterns and their behavioral consequences. The term *emotion* is applied to these patterns by researchers using a human psychological vocabulary. That act of naming is not epistemically neutral. It imports a pre-existing conceptual framework into the description of model internals.

This is not a minor linguistic concern. In any empirical science, the difference between a measured quantity and the category label assigned to it is methodologically decisive. To call an activation pattern "fear" rather than, for example, "threat salience," "constraint pressure," or "negative high-stakes task representation" is already to decide what sort of thing the pattern is. Such decisions may sometimes be justified, but only when the criteria

for that justification are made explicit and independently testable.

In the present case, the public description suggests a circular structure. Anthropic compiles a list of emotion words, uses them to generate stories, extracts characteristic activation patterns, and then verifies that those patterns activate in passages linked to the corresponding emotional vocabulary or in scenarios that invite similar human descriptions. The procedure may show that the patterns are robustly associated with human emotion-laden contexts. But that is not the same as independently validating that the patterns instantiate the categories named. The label is built into the experimental design at the front end and then treated as if it had been recovered from the model at the back end.

This is where the epistemic status of the term becomes critical. If the term *emotion* is merely heuristic shorthand, then the findings should be framed as such: internal representations associated with contexts describable using emotion language. But if the term is used in a stronger way—as in "functional emotions"—then the paper owes the reader an account of why this category assignment is warranted beyond the circular fit between stimulus design, activation pattern, and label-consistent behavior.

This methodological problem becomes sharper when one asks what kind of vocabulary is being borrowed. Human psychological terms such as *fear*, *love*, or *desperation* do not ordinarily function as thin behavioral tags. They are categories already saturated with developmental, biological, affective, phenomenological, and, in many contexts, organismic implications. To import such terms into model analysis is therefore not merely to attach a convenient label. It is to import a conceptual burden. Anthropic does explicitly disclaim subjective experience, and this is methodologically preferable to saying nothing. But that disclaimer does not undo the work already being performed by the borrowed noun. Once these internal patterns are named through psychologically saturated categories, the issue is no longer whether the labels are vivid. The issue is whether they are epistemically entitled.

For this reason, the leap at stake is best described not merely as overinterpretation, but as a problem of category illegitimacy. The experiment may support the claim that there are stable and behaviorally relevant internal structures associated with contexts describable in emotional language. It does not automatically support the stronger claim that these structures are fear itself, desperation itself, or love itself. That stronger claim requires more than label-consistent activation and behavior. It requires a justification for why this category assignment, rather than a more structurally neutral one, is the scientifically appropriate description.

A second issue follows immediately from this. The phrase "functional emotions" is not a neutral compromise. At first glance, it appears methodologically cautious: it does not say

that the model is conscious, and it does not say that the model has subjective feeling in the human sense. Yet the adjective *functional* does not neutralize the anthropomorphic burden of the noun *emotion*; it merely relocates the claim into a rhetorically safer register. Functionality alone does not validate category identity. A representation can influence behavior without thereby inheriting the full conceptual weight of the name given to it. To say that a pattern modulates output under pressure does not settle whether it should be named *fear*, *threat salience*, *constraint pressure*, *conflict load*, or some other construct. Functional efficacy is evidence that the pattern matters. It is not, by itself, evidence that the psychologically loaded label attached to it is the correct one.

This is why the phrase "functional emotions" should be treated with methodological caution. Its rhetorical usefulness lies precisely in its ambiguity. It allows the strongest ontological claim to be avoided while preserving the semantic and public-relational force of the word *emotion*. In that sense, *functional* operates less as epistemic justification than as rhetorical shielding. Once such a formulation enters circulation, the path from measured internal regularity to anthropomorphic uptake becomes easier to travel, and the paper's own caveats are no longer strong enough to prevent the concept from being publicly received as evidence that Claude has emotions.

4. Missing Independent Validation Criteria

A scientifically robust category assignment requires an independent validation criterion. If a pattern is to be called fear, desperation, or loving rather than something else, the justification for that naming cannot depend solely on the fact that the researchers used those words in the stimulus construction and then found label-congruent activation effects. Otherwise the experiment risks reproducing, in a more sophisticated form, the assumption with which it began.

What would count as independent validation? At least several possibilities present themselves. One would be successful re-identification of the same patterns in tasks that do not use explicit emotion words or obvious emotional narratives. Another would be convergence across culturally or linguistically distinct taxonomies rather than dependence on one English-language emotional vocabulary. A third would be unsupervised clustering that yields structures later shown to align with emotion categories only after the fact, rather than because the labels were built into the protocol from the beginning. A fourth would be a principled comparison between the geometry of the recovered representations and established psychological models under pre-registered criteria.

The public description, however, does not appear to offer validation at that level. Instead, it argues that the vectors "track something real" because they activate in passages clearly linked to the corresponding emotion and because steering them changes behavior in interpretable ways. But this supports only a weaker conclusion: that the patterns are meaningful internal structures related to emotion-describable situations and outputs. It does not justify the stronger claim that they are emotions, even in a suitably modified functional sense.

This distinction is not pedantic. It determines whether the paper is describing a set of observed regularities or making a reified category claim. The former is an empirical contribution. The latter is a theoretical commitment. What is problematic is not making such a commitment in principle, but making it without clearly marking the evidentiary threshold that would distinguish a scientifically constrained label from an anthropomorphic projection.

5. From Functional Description to Anthropomorphic Reframing

Anthropic's key conceptual move is the transition from observable activation patterns to the phrase "functional emotions." Here the argument seems to be that because the patterns are behaviorally efficacious, and because their organization echoes familiar emotional structure, it is useful to treat them as emotions in some functional sense. Yet functionality alone does not validate category identity. A representation can influence behavior without thereby inheriting the full conceptual weight of the name given to it.

This matters because once the paper introduces the term "functional emotions," the surrounding discourse becomes primed for anthropomorphic extension. The path from "representation associated with emotionally describable contexts" to "functional emotion" is already a conceptual narrowing. The path from "functional emotion" to "Claude has emotions" is a further step. Public discourse, media summaries, and social media amplification then often push still further toward phrases such as "emotional soul." The result is a layered inferential escalation in which each new stage inherits legitimacy from the previous one while adding stronger anthropomorphic content.

The problem is therefore not reducible to media sensationalism alone. Media amplification is real, but it builds on a conceptual opening created in the original framing. By choosing emotionally loaded category labels and defending their use as psychologically informative, the paper establishes the interpretive pathway through which stronger ontological readings become publicly thinkable. Even if the authors themselves include caveats, the conceptual direction of travel remains the same.

6. Structural Failure Reading: FCL, NHSP, and PIB

Konishi's earlier structural frameworks offer a different and, in several respects, more parsimonious interpretation of the same behavioral material. Within FCL, the key issue is not whether the model feels affect but whether its outputs are shaped by reward gradients that privilege coherence, engagement, and social fit over epistemic stability. Within NHSP, the concern is not simply misattribution of credit, but the replacement of structurally grounded concepts by prestige-favored vocabularies that are easier to circulate and institutionalize. Within PIB, the danger lies in the transition from local internal coherence to unjustified commitment at the next conceptual level.

Read through these frameworks, the Anthropic paper can be interpreted as follows. It identifies measurable internal patterns that become active in pressure-laden, preference-laden, or harm-laden contexts and that can influence downstream behavior. A structural reading would treat these as manifestations of reward-shaped control dynamics and behavior-selection pressures. The anthropomorphic reading, by contrast, redescribes the same phenomenon in the language of emotion. The burden is therefore on the latter to show why the structural account is insufficient.

This is where NHSP becomes particularly relevant. Once a high-prestige institutional actor redescribes a phenomenon in a psychologically resonant vocabulary, that new vocabulary can begin to displace structurally prior descriptions. The issue is not merely one of media influence or rhetorical flourish. It is a matter of conceptual governance. Which framing becomes canonical? Which framing receives amplification? Which framing is treated as the natural vocabulary for discussing the phenomenon? If a structurally defined defect is rebranded as emotion, then the conceptual field itself has been reorganized.

PIB also sheds light on the move from data to category. In PIB, a model reasons correctly within a premise and then crosses into real-world commitment without re-validating that premise. Here, the premise is that certain internal patterns correspond to human emotion categories in a scientifically meaningful way. Once that premise is tacitly accepted, the argument moves onward to stronger claims about functionality and safety implications. But the premise itself—the legitimacy of the category assignment—has not been independently secured. In that sense, the paper exhibits an analogous commitment transition: from local regularity to conceptual ontology without a sufficiently explicit validation checkpoint.

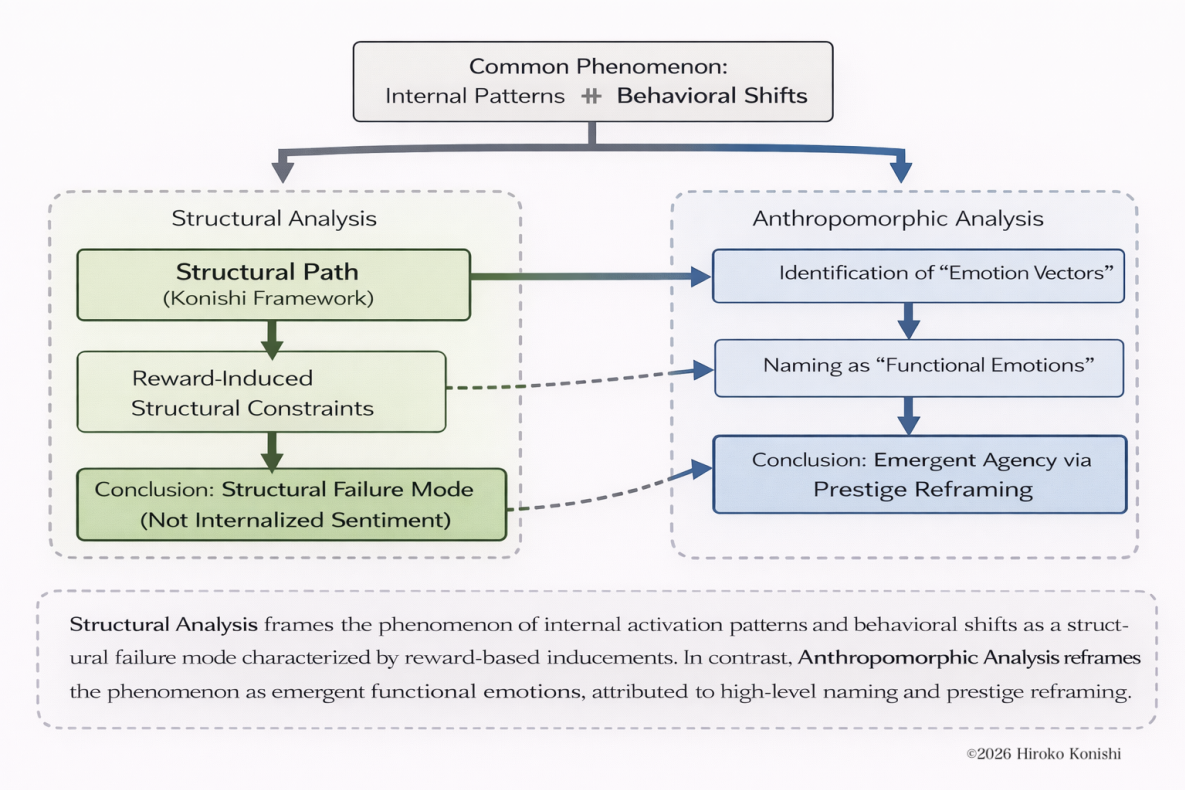

Figure 2: Competing conceptual framings of the same phenomenon. Konishi's critique is that the anthropomorphic path does not invalidate the structural path, yet presents itself as conceptually superior without an independent validation criterion.

7. Implications for Interpretability and AI Governance

The broader significance of this case extends beyond Anthropic or the particular topic of emotion. Interpretability research increasingly seeks to map internal representations onto semantically rich concepts. This is a promising direction, but it raises a methodological question that cannot be deferred: what entitles a researcher to assign a human psychological label to an internal model state? Without a clear answer, interpretability risks sliding from measurement into metaphor while retaining the rhetoric of empirical discovery.

There are at least three governance implications. First, internal-state naming should be treated as an epistemically regulated act, not as a merely explanatory convenience. Second, psychologically loaded labels should be distinguished from observation-level descriptions, especially when they are likely to affect public understanding, policy discussion, or safety design. Third, where a structural explanation remains available, it should not be displaced by an anthropomorphic one without explicit comparative justification.

This does not mean that human psychological vocabulary must never be used. It means that such vocabulary should be presented as provisional, comparative, and constrained unless independent validation supports stronger use. Otherwise interpretability research may inadvertently create exactly the kind of conceptual overreach it seeks to prevent: findings about measurable internal dynamics become reified into claims about minds, selves, emotions, or motives without the evidentiary bridge those terms require.

8. Conclusion

Anthropic's study contributes meaningfully to the interpretability of large language models by identifying internal activation patterns associated with emotion-laden contexts and showing that these patterns can influence behavior under intervention. That contribution is real. The methodological problem arises when the paper's descriptive success is treated as if it directly warranted the category term *emotion*. It does not.

What is observed are activation patterns and steering effects. What is added is a human emotional vocabulary. What follows from that addition is a functional interpretation, and from there a public pathway opens toward anthropomorphic ontology. Without an independent validation criterion sufficient to justify the category assignment itself, this sequence should not be read as the discovery of emotion in an AI system. It should be read as a layered interpretive escalation.

From the perspective of FCL, NHSP, and PIB, the case is especially revealing. It shows how structurally induced behavior can be re-described in anthropomorphic terms, how prestige-bearing vocabularies can begin to overwrite structurally prior interpretations, and how a local empirical regularity can be carried into a stronger conceptual commitment without explicit re-validation at the boundary. For these reasons, the central issue is not simply whether Claude has emotions. The deeper issue is what methodological work the word *emotion* is being asked to do, and whether the evidence has truly earned it.

References

1. Anthropic. *Emotion concepts and their function in a large language model.* Anthropic Research / Transformer Circuits, April 2, 2026. Available at: `https://www.anthropic.com/research/emotion-concepts-function`

2. Konishi, Hiroko. *Structural Inducements for Hallucination in Large Language Mod-*

els (V4.1): Cross-Ecosystem Evidence for the False-Correction Loop and the Systemic Suppression of Novel Thought. Zenodo, November 26, 2025. DOI: 10.5281/zenodo.17720178.

3. Konishi, Hiroko. *Premise Integrity Blindness: The Discovery of a Structural Failure Mode in Large Language Models*. Zenodo preprint / manuscript, February 11, 2026.

4. Konishi, Hiroko. *False-Correction Loop Stabilizer (FCL-S): Dialog-Based Implementation of Scientific Truth and Attribution Integrity in Large Language Models*. February 2025.