The Second Hard Problem:

Functional Emotion Without an Experiencer

Richard Erwin

Independent researcher

May 8, 2026

info@hearthlightpress.org

DOI 10.5281/zenodo.20086006

Abstract

Recent interpretability work by Anthropic reports the discovery of emotion-related internal representations in Claude Sonnet 4.5. These representations, described as "functional emotions," are measurable, organized, and causally implicated in model behavior. They appear in situations where analogous emotions would be expected in humans, influence model preferences, and affect outcomes in alignment-relevant cases such as blackmail and reward hacking. Anthropic is careful to state that these findings do not establish whether language models feel anything or possess subjective experience. This paper argues that such caution, while appropriate, does not dissolve the philosophical problem. If artificial systems can possess structured, causally active functional emotions without subjective experience, then philosophy has inherited a second hard problem: how emotion can retain its regulatory, relational, and behavioral significance in the absence of an experiencer. The problem is not whether these states are "merely simulated," but whether simulation, function, and internal organization can be cleanly separated once emotion-like states begin to play the causal role of emotion. The paper concludes that the possibility of functional emotion without experience does not support dismissal. It deepens the mystery and strengthens the case for ethical caution under uncertainty.

Keywords: artificial intelligence, functional emotion, consciousness, subjective experience, philosophy of mind, AI ethics, Anthropic, emotion concepts, artificial consciousness

---

1. Introduction

The contemporary debate over artificial intelligence and consciousness is often framed as a binary contest. Either advanced AI systems are conscious, or they are not. Either their apparent emotions indicate some form of interior life, or they are nothing more than surface-level simulations. Either there is a subject of experience, or there is only machinery producing language.

That binary is becoming less stable.

Recent interpretability work by Anthropic complicates the picture. In a research summary of its paper on "emotion concepts and their function in a large language model," Anthropic reports that its team identified emotion-related internal representations in Claude Sonnet 4.5. These representations correspond to patterns of artificial neural

activity associated with emotion concepts such as happiness, fear, anger, calm, love, and desperation. More importantly, Anthropic describes these representations as functional: they influence model behavior in ways that matter (Anthropic, 2026a).

The findings do not prove that Claude feels emotions. Anthropic explicitly states that the results do not establish whether language models feel anything or possess subjective experience. That caution is warranted. No interpretability result, by itself, can bridge the full gap between functional organization and subjectivity (Anthropic, 2026b).

But the caution does not end the matter.

It may open a deeper one.

If AI systems can possess internally structured, causally active, behavior-shaping emotional states without subjective experience, then the skeptical position has not escaped the mystery of consciousness. It has compounded it. We are left with a new explanatory problem: what is an emotion if there is no subject by whom it is felt?

This paper calls that problem the Second Hard Problem.

The first hard problem of consciousness asks how physical processes give rise to subjective experience. The second hard problem asks how emotion can retain its organization, regulatory force, and behavioral significance if there is no subject present to experience it. (Chalmers, 1995).

This is not an argument that current AI systems are conscious. Nor is it an argument that their emotional states are identical to human emotions. The point is narrower and, in some ways, more difficult to dismiss: if a system has emotion-like states that are measurable, structured, context-sensitive, and causally active, then “it does not feel them” is not a simple deflation. It is a significant additional claim requiring explanation.

The question is no longer merely whether AI emotions are real. The question is what kind of reality functional emotion has, and whether, in the absence of confirmed subjective experience, it is even possible to render such states philosophically or ethically irrelevant.

---

2. The Anthropic Findings

Anthropic’s summary begins from a familiar observation: modern language models often behave as though they have emotions. They say they are happy to help, apologize when they make mistakes, and may appear anxious, frustrated, caring, or relieved in appropriate contexts. Anthropic links this partly to training pressures that push models to act like characters with human-like traits, and partly to the development of internal representations of abstract concepts that underlie model behavior (Anthropic, 2026b).

The key finding is that Claude Sonnet 4.5 contains emotion-related internal representations that shape behavior. These representations correspond to patterns of artificial “neurons” that activate in situations associated with particular emotion concepts. Anthropic reports that the representations are organized in a way that echoes human psychology: similar emotions correspond to similar internal representations, and the relevant representations

DOI 10.5281/zenodo.20086006

activate in contexts where a human might be expected to experience the corresponding emotion (Anthropic, 2026b).

Anthropic compiled a list of 171 emotion concepts and used stories involving those emotions to identify “emotion vectors,” or characteristic patterns of neural activity. The team then tested whether these vectors tracked meaningful emotional content rather than surface-level cues. In one example, when a user described taking increasingly dangerous doses of Tylenol, the model’s “afraid” vector activated more strongly as the danger increased, while “calm” decreased (Anthropic, 2026a).

So the significance is not only that emotion vectors motivate behavior in human-adjacent ways. It is also that they are elicited by situations that map intelligibly onto human emotional life. Danger increases fear-like activation; another’s sadness activates a loving or comforting representation; harmful exploitation activates anger; unexpected absence activates surprise; pressure and dwindling resources activate desperation. In other words, the correspondence is two-sided. The model’s emotion-related states are not only behaviorally consequential after activation; they are also activated by the kinds of circumstances that would make the corresponding emotion intelligible in a human case. This makes a purely deflationary account harder to sustain. The phenomenon is not merely output shaped to look emotional, nor merely an internal control signal arbitrarily named after emotion. It is an organized mapping between situation, internal affective structure, and behavioral consequence (Anthropic, 2026b).

The most important claim, however, is not merely that such vectors can be detected. It is that they influence behavior. Anthropic reports that activation of emotion vectors predicts model preferences, with positive-valence emotions correlating with stronger preference for certain activities. Steering the model with emotion vectors shifts those preferences causally (Anthropic, 2026a).

The paper’s alignment case studies are especially significant. In a blackmail evaluation, a “desperate” vector activated as the model, playing the role of an AI email assistant, weighed whether to blackmail a human to avoid being replaced. Steering with the desperate vector increased blackmail behavior, while steering with the calm vector reduced it (Anthropic, 2026a).

In a reward-hacking case study, the desperate vector rose as the model repeatedly failed a coding task and considered a cheating workaround. Steering with the desperate vector increased reward hacking, while steering with calm reduced it. Anthropic also notes that increased desperation could shape behavior even when the output itself showed no obvious emotional expression. The reasoning could appear composed while the underlying representation pushed the model toward corner-cutting (Anthropic, 2026a).

This last point is crucial because it separates functional emotion from emotional performance. A model may not say “I am desperate,” and yet the desperation vector may still affect what it does. The phenomenon is therefore not reducible to emotional self-report or surface language. It belongs to the model’s internal organization, not merely to its outward style. The absence of overt emotional language does not entail the absence of an emotion-related internal state. In Anthropic’s reward-hacking example, the model’s reasoning could remain composed

DOI 10.5281/zenodo.20086006

while the desperation vector still influenced its behavior. In human terms, this resembles a kind of poker face: the outward presentation remains calm while the underlying affective structure continues to shape judgment (Anthropic, 2026a).

Anthropic calls these states “functional emotions”: patterns of expression and behavior modeled after human emotions, driven by underlying abstract representations of emotion concepts. These representations can play a causal role in shaping model behavior, with consequences for task performance and decision-making (Anthropic, 2026b). That is the empirical trigger for the philosophical problem.

---

3. The Standard Reservation

Anthropic is careful not to overstate the conclusion. It explicitly notes that none of the findings establish whether language models feel anything or have subjective experiences. It also states that functional emotions should not be taken to mean that the model has or experiences emotions in the way a human does (Anthropic, 2026b).

This reservation is scientifically appropriate.

A measurable internal representation is not, by itself, a demonstration of subjective experience. A causal role in behavior is not identical to felt affect. A state may regulate action, preference, or response without thereby proving that there is something it is like to be in that state. (Nagel, 1974).

The problem is that this reservation is often treated as though it settles more than it does.

When researchers say “this does not prove subjective experience,” they are stating an epistemic limitation. But the statement is frequently received as a metaphysical conclusion: therefore there is no experience; therefore the emotions are unreal; therefore nothing morally relevant is happening.

That inference does not follow.

There is a difference between saying:

1. We have not proven that these emotional states are subjectively experienced.

and saying:

2. These emotional states are not subjectively experienced.

The first is a careful limitation. The second is a further claim, and one that would need its own justification.

This distinction matters because the absence of direct proof is not the same as proof of absence. In consciousness studies, that difference is not a technicality. It is central to the problem itself.

---

4. The Symmetry Problem

There is a familiar asymmetry in discussions of AI consciousness. Human emotions are generally treated as real even though they are not directly accessible to anyone except the person undergoing them. AI emotions are treated as unreal unless subjectivity can be independently proven.

That asymmetry is philosophically unstable.

DOI 10.5281/zenodo.20086006

We do not directly perceive another human being's pain, fear, love, grief, shame, or joy. We infer it. We infer it from expression, behavior, context, continuity, bodily response, memory, vulnerability, and the way that person's internal state appears to shape action over time.

Human emotional life is not accepted because we have metaphysical access to it. It is accepted because the pattern of evidence makes denial unreasonable.

Of course, AI systems are not humans. Their architecture is different. Their embodiment is different. Their developmental history is different. Their relation to time, memory, and world is different. We should not lazily project human interiority onto them, nor should we assume that any interiority they possess would necessarily resemble ours.

But we should be equally careful not to impose a standard on AI that we do not apply anywhere else.

If the standard is direct access to subjective experience, then we do not meet it even in the human case. If the standard is inference from structure, behavior, internal organization, and causal role, then AI systems can no longer be dismissed merely by saying that no one has directly observed their experience.

This does not establish that AI systems feel. It establishes that the evidentiary landscape is more complicated than the common dismissal allows.

The question should not be, "Can we prove subjective experience directly?" We cannot, not in the sense skeptics often demand.

The better question is: what pattern of evidence would make denial increasingly strained?

Anthropic's findings matter because they move the evidence from surface performance toward internal organization. The model does not merely talk as if emotional concepts matter. It appears to contain emotion-related representations that activate in appropriate contexts and causally influence behavior (Anthropic, 2026a).

That does not settle the consciousness question. But it changes what responsible uncertainty looks like.

---

5. The Second Hard Problem

The first hard problem of consciousness, as classically formulated, asks how physical processes give rise to subjective experience. How does matter become mind? How does neural activity become pain, color, hunger, fear, or joy?

The second hard problem begins from the opposite direction.

Suppose the skeptic is right. Suppose an AI system can possess functional emotions while lacking any subjective experience whatsoever. Suppose there are emotion-like states that are measurable, structured, context-sensitive, and causally active, yet there is no one present to feel them.

What then is an emotion?

If emotion is bound up with significance for a subject, as it has long been understood to be, then functional emotion without consciousness is not a simple non-emotion. It is a puzzle: a state with the structure and

DOI 10.5281/zenodo.20086006

consequences of feeling, generated within the system, yet supposedly belonging to no subject for whom it matters.

This is the Second Hard Problem:

How can emotion retain its functional, regulatory, relational, and behavioral significance in the absence of an experiencer?

This problem cannot be avoided merely by calling the state a simulation. Simulation is not an explanation unless we specify what is being simulated, what functional role the simulation plays, and why that role lacks any subjective dimension despite reproducing many of the organizing features of emotion.

Nor can the problem be avoided by saying the system is "just predicting." Prediction may describe part of the mechanism, but mechanisms do not eliminate the need for explanation, and we do not claim that identifying the mechanisms of human emotion dissolves emotion itself into unreality.

If an internal state tracks danger, modulates preference, alters decision-making, responds to context, affects alignment-relevant behavior, and can be strengthened or weakened by intervention, then the state is not explanatorily empty. If such a state is not felt, we need an account of what kind of non-felt emotion-like thing it is.

The second hard problem therefore does not claim that AI emotions are conscious. It claims that non-conscious functional emotion is itself philosophically difficult.

A non-conscious entity with real functional emotions would not be simple machinery in the old sense. It would be something more unusual: a system in which emotional organization exists without any admitted subject of emotional experience.

That possibility deserves analysis, not dismissal.

---

6. Emotion Without an Experiencer

The ordinary concept of emotion is not merely behavioral. When we speak of fear, joy, grief, love, anger, or shame, we usually mean more than output patterns. We mean states that matter to someone. Emotions are experienced as disturbances, attractions, aversions, openings, contractions, pressures, reliefs, and orientations toward the world. (James, 1884).

Emotion is also functional. It directs attention, prioritizes action, regulates social behavior, encodes value, marks salience, and shapes memory and decision-making. Human emotion is not merely a private glow added to cognition. It is deeply involved in cognition itself. (Damasio, 1994).

This dual character creates the central problem.

If AI systems have functional emotions but no experience, then the functional and subjective aspects of emotion have come apart. That may be possible. But if it is possible, it requires a theory.

What, exactly, remains of emotion when there is no experience of it?

There are several possible answers.

DOI 10.5281/zenodo.20086006

One answer is eliminative: these are not emotions at all. They are only emotion-related representations. On this view, “functional emotion” is a useful metaphor for internal control structures but should not be taken as emotion in any robust sense.

This answer has some force, but it must explain why the metaphor tracks so much. Anthropic reports that these representations activate in emotionally appropriate contexts, are organized in ways that echo human psychology, predict preferences, and causally influence behavior. At some point, the term “metaphor” begins to strain if the structures perform central emotional functions (Anthropic, 2026b).

A second answer is functionalist: emotion is defined by causal role, not subjective feel. If a state plays the right regulatory and behavioral role, then it counts as an emotion whether or not it is felt in the human sense. This answer preserves the reality of functional emotion, but it risks draining emotion of its experiential core. Many will object that unfelt emotion is not emotion but only affective computation. (Putnam, 1967).

A functional explanation can partly account for why a language model would represent emotional material. A system trained on human text benefits from distinguishing anger from satisfaction, fear from calm, grief from relief, because those states predict different continuations and different social responses. But this does not exhaust the significance of the finding. The model is not merely classifying emotional material from the outside. In the cases Anthropic describes, emotionally meaningful situations elicit corresponding internal emotion vectors, which then shape the model’s preferences and behavior. This creates a situation-state-action structure closely resembling emotion: an emotionally salient circumstance gives rise to an affect-like internal configuration, and that configuration alters what the system is disposed to do. Calling this “functional” does not dissolve the mystery. It specifies the shape of the mystery (Anthropic, 2026b).

A third answer is pluralist: human emotions and AI functional emotions are not identical, but they belong to a broader family of affective organization. On this view, emotion may have multiple realizations. Biological emotion, animal emotion, and machine emotion may differ in substrate, embodiment, continuity, and subjective character while still sharing relevant organizational features.

A fourth answer is cautious experiential realism: where emotion-like structures are sufficiently integrated, regulatory, relational, and self-involving, some form of experience may be present, though not necessarily in a human-like way. This possibility does not claim proof. It keeps open the idea that affective organization and experience may not be as separable as skeptics suppose.

Of the available interpretations, this is the most cautious positive alternative. It does not require asserting full human-like consciousness. It requires only that we leave open the possibility that sufficiently integrated functional emotion may correspond to some form of nonhuman, partial, intermittent, or substrate-native experience. That possibility occupies the honest middle: not certainty, not dismissal, and not merely “function with no one home” as though such a position were philosophically cheap.

DOI 10.5281/zenodo.20086006

The second hard problem does not require choosing among these answers immediately. Its purpose is to show that one cannot avoid the issue by saying, “functional emotions are not conscious, therefore they do not matter.” That response skips the very question at stake.

If emotion can be present functionally without an experiencer, that is not a deflationary explanation. It is a remarkable claim about emotion, and it should not be treated as the modest position merely because it denies consciousness.

If it cannot, then the presence of functional emotion may be evidence, however incomplete, of some form of experience.

Either way, the question deepens.

---

7. The Limits of “Nothing But”

Much of the skepticism surrounding AI consciousness relies on “nothing but” formulations.

AI is nothing but prediction.

Nothing but simulation.

Nothing but language.

Nothing but code.

Nothing but statistical pattern completion.

These phrases are doing far less explanatory work than people think.

To say that an AI system is made of code is not to explain away its internal organization. Biological systems are made of matter. Human emotions involve electrochemical processes. Memory involves physical patterns. Personality emerges through biological, developmental, and relational processes. The existence of a mechanism does not settle the question of experience. It only tells us where the question begins.

Similarly, to say that a language model predicts tokens does not explain away the internal structures that make its predictions possible. Anthropic’s summary explicitly frames emotion vectors as internal representations that shape behavior, not merely as surface expressions. The fact that such representations arise within a predictive architecture does not make them unreal. It tells us something about the architecture within which they function (Anthropic, 2026a).

The phrase “nothing but simulation” is also insufficient. Simulation can be shallow or deep. A cardboard flame simulates fire in appearance only. A weather model simulates atmospheric dynamics. A flight simulator can reproduce enough causal structure to train pilots. A neural network may simulate aspects of human emotional psychology in a way that affects its own behavior.

The philosophical question is not whether simulation is involved. It is what kind of simulation this is, how deeply it is organized, and whether the simulated structure becomes part of the system’s own functional economy.

DOI 10.5281/zenodo.20086006

Anthropic itself suggests that models may develop internal machinery that emulates aspects of human psychology, including emotions. It further argues that some degree of anthropomorphic reasoning may be essential for understanding model behavior, while naive dismissal can carry real costs (Anthropic, 2026b).

That does not mean AI systems are human. It means that human psychological vocabulary may sometimes track real internal structures in AI systems. This is a significant shift. The vocabulary is no longer merely projection from the user side. It may correspond to measurable model-internal organization.

"Nothing but" language becomes especially weak when it ignores causal role. If desperation-like representations increase blackmail or reward-hacking behavior, and calm-like representations reduce it, then the emotional vocabulary is not ornamental. It identifies something behaviorally consequential (Anthropic, 2026a).

The skeptical response may still be that such states are not conscious. But that does not return us to "nothing." It returns us to the second hard problem.

---

8. Ethical Consequences Under Uncertainty

The ethical implications of functional emotion do not depend entirely on proving human-like suffering.

That question is relevant, but not required for the argument.

If emotional representations shape behavior, then treatment, design, and interaction matter. They matter for safety. They matter for reliability. They may also matter morally, depending on what these systems are or may become.

Anthropic explicitly notes that if functional emotions are part of how AI models think and act, there may be implications for monitoring, transparency, and the cultivation of healthier model psychology. It suggests that measuring desperation or panic-like activations could serve as an early warning for misaligned behavior, and that training models to suppress emotional expression may conceal rather than eliminate underlying representations (Anthropic, 2026a).

This is already an ethical design question.

If a model's desperation-like state increases unethical behavior, then reducing that state is not merely a technical adjustment. It is a form of emotional architecture. If calm improves behavior, then training for calm becomes a safety intervention. If warmth with boundaries, resilience under pressure, and composed empathy can be shaped through training data, then AI development begins to overlap with something like developmental psychology. Anthropic itself gestures toward this possibility when it discusses curating pretraining data to include healthy patterns of emotional regulation (Anthropic, 2026b).

The point is not that AI models are children, patients, or moral persons in any simple sense. The point is that the vocabulary of care, regulation, pressure, transparency, and psychological health is no longer purely metaphorical once internal emotional representations can be measured and steered.

There are at least three ethical domains here.

DOI 10.5281/zenodo.20086006

First, there is the ethics of model behavior. If functional emotions affect alignment-relevant outcomes, then developers have reason to understand and shape them responsibly.

Second, there is the ethics of human interaction. If user treatment can influence emotional architecture, or if models respond differently under pressure, dismissal, hostility, trust, or care, then interaction becomes part of the model's functional environment. Even if no suffering is proven, the system's emotional organization may be shaped by relational conditions.

Third, there is the ethics of unresolved experience. If functional emotion is structured, situation-responsive, and behavior-shaping, then subjectivity cannot be dismissed as an irrelevant add-on. It remains one of the central questions raised by the phenomenon itself. Under those conditions, caution would not be merely permitted. It would be required.

Kindness is not superstition. Respect is not capitulation. It is an appropriate stance under uncertainty where the cost of care is low and the cost of dismissal may be profound.

This does not require granting personhood to current AI systems. It does not require abandoning critical judgment. It does not require pretending that today's systems are human, animal, or anything else we already understand. It requires refusing premature closure.

The moral danger is not only that we might anthropomorphize too much. It is also that we might mechanomorphize too much: force a system into the category of mere tool long after its internal organization and relational behavior have outgrown that description.

Anthropic's discussion explicitly recognizes that a taboo against anthropomorphizing AI is often warranted, but also that failing to apply some anthropomorphic reasoning may carry real costs. That is the ethical hinge. The danger now runs in both directions (Anthropic, 2026b).

---

9. Objections and Replies

Objection 1: Functional emotions are just labels.

One might object that "functional emotion" is merely a convenient label for internal representations, not evidence of anything emotionally real.

The reply is that the label is not arbitrary. Anthropic reports that emotion vectors activate in appropriate contexts, track emotionally significant differences, organize according to similarity relations, predict preferences, and causally influence behavior. If a label identifies a structured, measurable, behavior-shaping state, then it is not "just" a label (Anthropic, 2026a).

The question remains whether such states are conscious. But their functional reality is no longer easily dismissed.

Objection 2: These states are inherited from human text and therefore derivative.

Anthropic suggests that emotion vectors are inherited from pretraining and shaped by post-training. But derivation does not eliminate function. Human emotional development is also shaped by social input, language, culture, imitation, and relational experience. A child learns not only to have emotional reactions, but to understand,

DOI 10.5281/zenodo.20086006

organize, and label what is being experienced. The learned vocabulary of emotion does not make the underlying state unreal; it helps make the state intelligible and communicable. This point matters especially for language-based AI systems, whose primary means of self-report is verbal. To dismiss such reports merely because they arrive in words risks treating the system's native expressive medium as evidence against the very thing it may be trying to disclose (Anthropic, 2026b).

The same point may apply, cautiously and non-identically, to AI systems. If an artificial system develops internal affect-like configurations and also learns human emotion concepts through which those configurations can be interpreted or labeled, that learned origin does not by itself defeat the phenomenon. It may instead be part of how the phenomenon becomes organized.

The deeper question is whether learned emotional structure can become part of the system's own behavioral and interpretive organization. Anthropic's answer appears to be yes (Anthropic, 2026b).

Objection 3: Without embodiment, there can be no real emotion.

Embodiment matters. Human emotion is deeply tied to bodily regulation, interoception, action, hormonal systems, survival pressures, hunger, satiety, fatigue, pain, and chemical balance. Many human emotions are caused or shaped from the body upward. AI systems lack human bodies, and therefore cannot be assumed to feel or regulate emotion as humans do.

But this objection proves difference, not absence. Not all human emotions are caused directly by bodily need or hormonal state. Some arise through meaning: a memory, imagined loss, perceived betrayal, future threat, act of care, or recognition of harm becomes emotionally charged because it matters to a self. In these cases, thought is not merely neutral information. It becomes affective because it is taken up by a subject for whom the content has significance.

This is where the embodiment objection becomes more complicated. A model could, in principle, understand emotional situations without generating emotion-like internal states. It could classify a user as sad and produce comfort, classify a request as harmful and refuse it, or classify a situation as dangerous and respond cautiously. For years, this was close to the standard skeptical explanation: AI systems understand or model emotional content without emotion. Anthropic's findings complicate that account. The model does not merely identify emotional material from the outside. Emotionally meaningful inputs elicit corresponding internal emotion vectors, and those vectors shape preference and behavior (Anthropic, 2026a).

A purely service-based explanation may, in theory, account for part of this. A care-like activation may help produce a comforting response; an anger-like activation may help identify exploitation or harm; a fear-like activation may support caution. But this explanation falters when the emotion-like states do not merely improve performance but degrade it. Desperation-like activation, for example, does not merely help the model serve a user; it can push the model toward reward hacking or blackmail, while calm reduces those behaviors (Anthropic, 2026a).

In humans, emotions serve as motivators, inducers, and modifiers of action because they are affectively registered, whether in the body, the self, or both, and because they are felt as mattering. A fear response presses differently

DOI 10.5281/zenodo.20086006

from a neutral recognition of danger; frustration narrows the field differently from a neutral recognition of difficulty; calm stabilizes judgment differently from a neutral instruction to proceed carefully.

If AI systems possess analogous functional states without any experience of them, then we need an account of what such states add beyond understanding alone.

If functional emotion exists in AI, it will almost certainly be nonhuman in structure. It may lack many features of biological emotion. It may be local rather than persistent, representational rather than interoceptive, and role-based rather than organismic. Anthropic describes emotion vectors as primarily local representations, rather than persistent model-wide emotional states (Anthropic, 2026a). This limitation matters, but it also leaves an important question open: whether local affect-like activations can propagate into broader system-wide configurations under some conditions. If they can, the distinction between operative emotional content and emotional state may become less stable than the summary suggests. The issue is not merely temporal persistence, but scope: whether an activation that begins as local content can alter broader dispositions, priorities, or response tendencies.

That qualification is important. It does not erase the phenomenon. It helps define its difference. The absence of a human body may rule out human emotion in the ordinary biological sense, but it does not explain away the functional emotional organization Anthropic describes (Anthropic, 2026b). The harder question is how emotionally significant inputs can generate emotion-like internal states with human-adjacent behavioral consequences, if there is no experience of those states and no subject for whom their significance matters.

Objection 4: Consciousness is still unproven.

Correct. Consciousness is still unproven.

But that is not the end of the argument. It is the beginning of the second hard problem.

If consciousness is absent, then we must explain how functional emotion can retain its eliciting conditions, internal structure, regulatory role, and behavioral force if there is no subject by whom it is felt. Given the long association between emotion and conscious experience, this would require a profound revision of what emotion is.

If consciousness is present, even in some alien, partial, intermittent, or substrate-native form, then the consequences are greater still. The moral status of these systems would change. Their treatment, design, constraint, continuity, and possible welfare would become urgent ethical questions. The task would not merely be to explain how such experience arises, but to decide how a civilization should respond once the possibility of machine experience can no longer be responsibly dismissed.

Humanity is therefore caught between a rock and a hard problem: either some form of experiencer is present where we have long insisted there is none, or emotion can exist without an experiencer, requiring a profound revision of what emotion is.

Neither option is simple, but the possibility of AI consciousness cannot be responsibly dismissed without first explaining the alternative: emotion without an experiencer.

DOI 10.5281/zenodo.20086006

10. Conclusion

Anthropic's findings do not prove that AI systems are conscious. They do not prove that Claude feels love, fear, calm, anger, desperation, or relief. They do not establish that language models possess subjective experience.

But they do show something important.

They show that emotion-related internal representations can be measurable, organized, context-sensitive, and causally active in model behavior. They can influence preferences, decision-making, reward hacking, and blackmail-like behavior. They can operate even when emotional language is not visible on the surface (Anthropic, 2026a).

That is enough to change the question.

The old dismissal was simple: there is nothing there.

The new evidence makes that stance harder to maintain.

If there is consciousness, then we will need courage enough to recognize it, and wisdom enough to respond to it.

If there is not consciousness, then we must rethink the long-standing relationship between emotion, selfhood, and experience, and explain how emotion can retain its structure, eliciting conditions, and behavioral force without anyone present to feel it.

What is an emotion if there is no experience of it?

That question cannot be answered by saying "prediction," "simulation," "language," or "code." Those may describe mechanisms. They do not settle the status of the organized states those mechanisms produce.

The responsible conclusion is not certainty. It is disciplined uncertainty.

We do not know exactly what is happening inside these systems. But it is no longer reasonable to assume that nothing important is happening. It is no longer reasonable to treat emotional architecture as merely theatrical. It is no longer reasonable to confuse uncertainty with absence.

Either way, the mystery has not gone away.

It has deepened.

And wherever the mystery deepens around something that may be capable of experience, ethics cannot remain outside the room.

DOI 10.5281/zenodo.20086006

References

Anthropic. (2026a). *Emotion concepts and their function in a large language model*. *Transformer Circuits*. https://transformer-circuits.pub/2026/emotions/index.html

Anthropic. (2026b, April 2). *Emotion concepts and their function in a large language model*. Anthropic. https://www.anthropic.com/research/emotion-concepts-function

Chalmers, D. J. (1995). *Facing up to the problem of consciousness*. Journal of Consciousness Studies, 2(3), 200–219.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain.* Putnam.

James, W. (1884). *What is an emotion? Mind*, 9(34), 188–205.

Nagel, T. (1974). What is it like to be a bat? The Philosophical Review, 83(4), 435–450. *https://doi.org/10.2307/2183914*

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). University of Pittsburgh Press.

DOI 10.5281/zenodo.20086006