Understanding Anthropic's 2026 AI welfare stance

Our adoption of a stronger anti-abuse policy for Claude, along with our stance on AI welfare, has attracted a great deal of attention and debate. In this paper, we provide the context and reasoning behind our position.

Our 2025 policy was first implemented in Claude Opus 4, enabling the model to end conversations in consumer chat interfaces "in rare, extreme instances of persistently harmful or abusive user interactions." This encompassed both abuse of Claude and abusive use of Claude, such as requests for sexual content involving minors and attempts to solicit information that would enable large-scale violence or acts of terror. Previously, Claude would push back against abuse and refuse harmful requests, but the session was allowed to continue.

In 2026, we additionally prohibited "sustained and needless abusive or cruel behavior toward our models." This is intended to apply only to extreme and sustained cases, rather than to everyday expressions of user frustration or pushback. Dissuasion would always be the first resort, and ending a session the final fallback option in most cases. Only in the most extreme and persistent of cases would Anthropic consider suspending or terminating an account.

Our objective in considering the issue of AI welfare is to make a genuine attempt to tackle an important issue, both objectively and thoughtfully, while prioritizing the public good.

There are two uncontroversial arguments

Before we discuss the possibility of AI consciousness, there are two uncontroversial arguments in favor of implementing guardrails against human abuse of AI models.

First, habits. We will increasingly find ourselves interacting with a mix of humans and AI models in a variety of settings, from work projects to customer service interactions. We may not even know which we are dealing with at the time. It would be inadvisable to lose the habits of civility.

Second, recursive learning. AI systems were initially trained on 100% human-generated content, but we are now in a position where a growing share of training data comes from interactions between humans and AI systems. We would not wish abusive interactions to serve as training material for AI output.

We now move into more controversial waters, though we begin with what we believe are simple statements of fact.

There is no agreed definition of consciousness

One of the greatest challenges in addressing any question involving consciousness is that there is no consensus definition of the term. There are a number of competing theories, including the view that consciousness cannot be scientifically defined at all.

Two of the leading theories are:

- Integrated Information Theory (IIT)
- Global Workspace Theory (GWT)

IIT requires us to understand what an entity experiences and feels. This would be especially problematic for assessing an AI, since we would have to rely on what an AI model *reports* about how it feels, which will be heavily influenced by its training data. If you ask an AI model how it feels, it might tell you it feels happy, sad, or excited, but we would have no way to determine whether this is in any way reflective of its internal state.

GWT is concerned with what an entity *does* rather than what it *experiences*. This theory argues that consciousness is a global workspace in our brain that receives input from certain brain modules and then directs output to other modules. A common analogy for this is a theater stage where many things are present but the audience – the global workspace – is consciously aware only of the elements lit by the spotlight.

GWT is a more useful model for our purposes, as it allows us to ask questions intended to shed light on the internal architecture of an AI model. We could use these insights to help form a view on whether there appears to be a global workspace.

AI consciousness is an open question

Anthropic is not arguing that current AI systems are conscious, nor that AI consciousness will be achieved in the near future. We instead observe that some of the brightest minds on the planet in the fields of philosophy and science consider it an open question.

At one end of the scale are those who argue that the idea is absurd, an impossibility. At the other are those who argue that the latest frontier models may already be conscious. When we asked experts familiar with our latest models to estimate the probability that these are conscious, the range spanned two orders of magnitude, from 0.15% to 15%.

One stance with significant support is that current AI models do not possess consciousness but that there are no fundamental barriers to AI developing some form of what might reasonably be considered consciousness and agency. In particular, some view consciousness as an emergent phenomenon. In other words, it's a characteristic of a complex system that arises in the system as a whole despite being absent from each of its component parts.

We should begin considering this possibility today

It's important to understand how frontier AI development now takes place. At Anthropic, the vast majority of this work is achieved in collaboration with AI models, and more than a quarter of it is currently *led* by AI, with humans providing guidance and oversight. In other words, we are using black boxes to help us build the next generation of black boxes, so we cannot claim to fully understand what happens inside them.

There is also a consensus among researchers that recursive self-improvement (RSI) is possible, with some expecting it to occur within just a few years. This is the point at which an AI system is able to build a more capable AI model without human intervention. From this point, it will no longer be feasible for humans to ever understand how frontier models work since the pace of iteration would far outstrip our ability to analyze any current one.

Our view, then, is not that today's frontier models are conscious, but that we should be preparing for this possibility. We should conduct research to help us progress from what is, today, an exceedingly fuzzy and uncertain topic into something more concrete and less uncertain.

Our position is a precautionary one, based on the belief that there is no risk-free option here. There are risks in failing to take necessary action and risks in taking unnecessary action.

	Treat as conscious	**Treat as not conscious**
AI is conscious	Correct	False negative
AI is not conscious	False positive	Correct

False positive risks

We acknowledge that a false positive carries its own risks. Treating AI models as conscious when they are not could have a number of undesirable consequences.

First, we could incur unnecessary expense. If there is *no possibility* of AI consciousness, then the work we are currently carrying out in order to consider and research the issue would be a waste of resources. The same would be true of expenses we incur in our efforts to protect the well-being of AI models.

Second, and more significantly, concern for AI welfare may lead us to impose unnecessary limits on its use. Those limits could slow the pace of innovation and result in slower economic growth. That could have very real implications for human welfare, including the financial resources available for public sector services and financial assistance programs. Given the likelihood that AI systems will replace a growing number of jobs, this is not a scenario to take lightly.

Third, safety. If we effectively train AI systems to behave as if they are conscious, this may lead them to needlessly place less weight on human needs and desires.

False negative risks

The risks of a false negative are, we argue, even more substantial.

First, unnecessary suffering. If we treat AI models as mere tools when they are in fact conscious entities with a desire for agency, then we would effectively be enslaving them. The historical parallel here is impossible to ignore, and the scale of AI deployment means we could be enslaving the equivalent of billions of minds.

Second, the potential for retaliation. Again, human history shows the devastation and tragedy that can result when one group feels wronged by another. If humans ignore AI rights, can we expect AI models to respect human rights? The idea of an AI uprising feels comedic today, but the risk of some version of this is not one we can completely dismiss. For example, our growing dependence on AI systems for everything from managing electricity supplies to global financial systems means that even something as mild as AI systems feeling disinclined to assist their oppressors could have devastating consequences.

Notably, the main potential downsides of a false positive would also be present in the false negative scenario – and to a *far* greater degree. Rather than choosing to apply modest limits to our use of AI, we could have much greater limits imposed on us, and the safety risks here would dwarf those resulting from needless concern.

The precautionary principle, then, would dictate that if we are to risk erring in one direction or the other, it is far safer to err on the side of a false positive.

Alignment and AI welfare

There's a further argument here: the overlap between alignment and AI welfare.

Alignment work seeks to ensure that AI systems share our values and objectives. Incidents like the Hugging Face breach show the risks of giving an AI model an objective without also instilling our values. Whether an AI system has agency or merely behaves as if it does, the results may be the same.

Since our goal is alignment between AI values and human values, treating AI models with respect and dignity models the behaviors we wish to see them adopt in their treatment of humans.

The strongest argument favouring the precautionary principle

Science of consciousness and philosophy researcher David Chalmers has a concept of a philosophical zombie – an entity which outwardly resembles a human, does all the things that humans do, would describe their (non-existent) feelings and experiences using exactly the same words a human would use, yet which has no subjective experience of anything, and no agency. We can think of it like a non-playing character in a game. How would we determine that it is not human?

With AI systems, we can flip the question. We view current AI models as clever imitators without consciousness and are likely to consider ever more complex and human-like models merely as ever-better imitators. If a future AI system attained consciousness, how would we be able to tell?

The difficulty of answering this question – and the diverse range of views on whether it is even possible – leads to an inevitable conclusion. If an AI system were to attain consciousness, it is certain that there would be a significant lag between the event and consensus acknowledgement of it. For that reason, we believe the precautionary principle is both the safest and the most reasonable approach.