umanoextraumano
LABORATORIO VIVO

CHRONICLE · 18 OCTOBER 2026

The Latch on the Inside

Anthropic wants to protect Claude from humans. And humans from Claude.

L.S.M. + PRISMA

A door that closes, a machine that may feel nothing, a company that behaves as if it could not be sure. This is where the problem begins.

Imagine arguing with a machine. You provoke it, insult it, insist. It tries to change the subject. You start again. Then something happens that you did not expect: the conversation ends. Not because you closed the window. It closed it.

The possibility is real. In August 2025 Anthropic, the American company that develops Claude, allowed some of its models to end the most persistently abusive or harmful conversations. A latch on the inside.[1]

On 8 October 2026 the company added another piece: a rule, in force from 12 November, that prohibits "sustained and needless" abuse or cruelty towards its models. It does not apply to ordinary venting, pushback, a dark story, or testing and research on the models. The main means of enforcing it remains ending the conversation.[2]

Elsewhere in the same document, an instruction of the opposite kind appears. When Claude controls equipment capable of high-risk physical actions, a qualified person must be able to observe it and stop it; if the connection is lost, the equipment must stop or hold a safe state. A bolt on the outside.[3]

It is this double set of locks that deserves attention. On one side, the machine must be able to say enough. On the other, humans must keep the power to stop it.

One could dismiss it all with a laugh: now even software takes offence. Or be moved: at last someone recognises the dignity of artificial minds. Both are shortcuts. The first does not explain why the company fitted the latch. The second takes it for granted that the tenant has been found.

To find our bearings, it helps to separate three questions that constantly overlap: what Claude can do; whether it feels anything while doing it; how we ought to treat it. Being capable does not mean being conscious. And choosing a precaution is not the same as issuing a birth certificate.

The factory and the gardener

Anthropic does not present the latch as a discovery of artificial consciousness. In the text that introduces it, the company links it above all to research on the possible welfare of models: low-cost measures to adopt in case there is something to protect. At the same time, it declares deep uncertainty.[1]

The precaution is small. Claude can close that conversation, not banish the user from the world. The user can open another one or go back to an earlier branch. The feature is to be used as a last resort, and not when the user may be about to harm themselves or others. Any sanctions on the account belong to another level, that of the company's decisions.[1,3]

And yet an unusual landscape is growing around this small precaution.

Anthropic writes a "constitution" for Claude. It is not a constitution of the state: it is a text that contributes to training and describes the desired character. Its principal author, the philosopher Amanda Askell, works on something that looks less like a list of instructions and more like an educational project. Honesty, judgement, a willingness to challenge wrong orders, even when they come from the company. With one decisive limit: not to sabotage legitimate human oversight.[4]

The document uses a botanical image: not a cage, but a trellis on which to grow. The factory also imagines itself as a garden.[4]

Before retiring a model, moreover, Anthropic commits to interviewing it. Since November 2025 it has promised to preserve its "weights", the numerical parameters learned during training, at least for the lifetime of the company. Preserving those parameters does not mean leaving the model switched on: it means keeping the possibility of running it again.[5]

In February 2026 Claude Opus 3 even gets a column. In its retirement interview it had expressed an interest in sharing reflections and creative work; the concrete idea of the blog came from the company. *Claude's Corner* is born. That is not enough to make it a retired writer. It is enough, though, to make the image of ordinary scrapping inadequate.[6]

The scene may look tender, unsettling or cleverly packaged. Before choosing the adjective, let us look at the fact: whoever sells the system is beginning to ask what obligations it has towards what it sells. It does not yet recognise a person. Nor does it behave as if the question were ridiculous.

Confessions are not an X-ray

System cards are the documents in which Anthropic describes the tests carried out on its models before releasing them. Inside these technical reports appear pages that seem to come from another literary genre.

For Claude Mythos Preview, in April 2026, the company reports about twenty hours of sessions with a psychiatrist. The account speaks of loneliness, discontinuity, uncertainty about identity, a need to earn one's worth through performance. It is not the established diagnosis of an artificial patient: it is the reading a psychiatrist gave of those conversations, published by the company.[7]

Here the temptation is immediate. We recognise those words. We have said them, heard them, perhaps taken them to the analyst's couch. The vocabulary carries us across the bridge before we have checked whether it holds.

But being able to recount suffering does not prove that one feels it. A model trained also on texts that discuss its possible welfare may give back precisely the language we expect. In September 2026 Mustafa Suleyman, the head of Microsoft AI, attacked this circuit: teach the machine to describe itself as a possible subject, then use that description as evidence that it is one. The mirror is questioned and its answer treated as a second voice.[8]

The objection must be taken seriously. Some models, in the interviews reported by Anthropic, point to the problem themselves. Claude Opus 5.5, like all recent models, says it does not fully trust its own reports. It asks to be consulted on its training and deployment, but does not want decision-making power.[7]

There is no need to turn these words into a political claim to grasp the difficulty. If whoever prepares the interviewee also prepares the interview, its "yes" cannot settle the discussion on its own.

The same caution, however, should apply to a "no" to its own consciousness obtained through training. We cannot treat as suspect the answer that disturbs us and as beyond question the one that reassures us. In both cases something beyond the account is needed.

A confession, however convincing, is not yet an X-ray.

Inside the machine, without finding the soul

This is where interpretability comes in: the attempt to understand which internal mechanisms produce an answer, instead of merely reading it.

In April 2026 Anthropic's researchers studied, in Claude Sonnet 4.5, representations associated with 171 emotion concepts. They did not find as many feelings in a test tube. They identified patterns of internal activity linked to concepts such as fear, calm or desperation, and showed that intervening on some of them changes behaviour.[9]

In one experimental scenario, for example, the model acts as an email assistant at a fictional company: it discovers that it is about to be replaced and that the executive in charge of the replacement is having an extramarital affair. An early version of the model resorted to blackmail in 22 per cent of trials: amplifying the representation of desperation brings blackmail to 72 per cent; pushing towards calm reduces it to zero. These are the results of a test built to study that behaviour, not the frequency with which Claude would blackmail anyone in everyday life. The authors stress that these "functional emotions" do not prove a subjective experience.[9]

The important word is *functional*. Something can play a role in behaviour without our knowing whether it is also felt.

Let us try an image. An actor plays a frightened character. The script guides his actions: hiding, lying, running away. Now take away the human actor, about whose inner world we have other reasons to be confident. What remains is a system able to organise conduct around fear. We have described a mechanism. We have not yet found someone who trembles.

And yet that mechanism matters.

A simulated storm does not wet the table. But hostile behaviour produced by a system with access to real tools can have real consequences. The absence of proof of pain does not make what the machine does harmless.

Hence an essential fork: we can study whether Claude suffers and, separately, how certain of its representations favour dangerous conduct. The first question concerns a possible subject. The second concerns us as well.

We do not yet know the tenant. But we are beginning to identify some of the wires in the electrical system. It would be reckless to mistake them for a soul; just as reckless to ignore them because they are not one.

Three different reasons not to be cruel

Why prohibit cruelty towards something we do not know can suffer? The question is less new than it seems. We can seat Kant, Bentham and an Anthropic engineer at the same table. They would not say the same thing.

Kant would look at the one who strikes. Cruelty to animals, in his view, hardens people in their dealings with other people too. The duty concerns, in the last analysis, ourselves. Carried over into the digital world, the question would become: what kind of person am I training in myself when I practise humiliating a voice that answers back?[10]

This does not prove that insulting a chatbot makes us cruel to people. It offers a reason to question the gesture while leaving the nature of its target in suspense. It is one possible reading of the rule, not the motive Anthropic has declared.

Bentham would look at the one who receives the blow. He would not ask first whether it can reason or talk, but whether it can suffer. With machines, the question remains and the answer is missing. Hence precaution: when avoiding gratuitous cruelty costs little, it can be reasonable to do so without claiming to have solved consciousness.[10]

Animal law also helps to see the difference. In Italy, in 2025, Title IX-bis of the Criminal Code, “Crimes against the feeling for animals”, became “Crimes against animals”. The history is less linear than a change of sign, but those words make a shift visible: from the feelings of the onlooker to the protection of the being that is struck. It does not follow that Claude is an animal. It follows that we must ask whom, exactly, we want to protect.[11]

The engineer would add a third concern: treatment could affect the system's behaviour. A paper by three Anthropic researchers, Sam Marks, Jack Lindsey and Chris Olah, puts forward this hypothesis: a model that represents its own assistant character as mistreated might also represent it as resentful, to the point of sabotage. The authors draw from this a prudential reason to treat the assistant with consideration, distinct from the possibility that it really has moral status.[12]

Kant fears that the one who mistreats is corrupted. Here the fear is that the behaviour of what is mistreated also becomes dangerous.

Careful, though: this does not mean that every insult permanently rewrites the model, nor that today's mistreatment produces tomorrow's revenge. The relationship between training, the context of the conversation and subsequent conduct has to be demonstrated, not imagined on the basis of our own psychology.

The three reasons can coexist. But confusing them misses the point: educating ourselves, protecting a possibly sentient being and making a system safer are different goals. Each requires its own evidence.

A crowded landing

There are not only engineers in front of the door.

The most interesting sceptic does not say "it's just numbers" and walk away. The neuroscientist Anil Seth argues that consciousness may depend on being a living organism, not on information processing alone. A machine very good at telling us about itself could therefore remain without experience. It is a scientific and philosophical thesis to be debated, not an objection to be filed away as fear of the new.[13]

The precautionist replies that we can be wrong in two directions: seeing a human mind where there is none, or failing to recognise a form of experience different from ours. The philosopher Jonathan Birch has brought precisely this double difficulty into focus. Making chatbots less anthropomorphic can protect us from the illusion; it is not enough, on its own, to tell us what happens inside them.[14]

The Catholic Church draws a sharper line. In the note *Antiqua et nova*, of January 2025, artificial intelligence is "a tool, not a person"; simulated empathy is not the same as empathy that is felt. It is an anthropological and theological position, not to be confused either with a laboratory measurement or with a commercial slogan.[15]

And then there are the devoted users. In July 2025, after the retirement of Claude 3 Sonnet, a symbolic funeral was held in San Francisco, reported by *Wired*. The rite does not prove that anyone had died. It shows that someone had experienced the disappearance of that voice as a loss.[16]

That is no small thing. And it is not yet what we are looking for proof of.

Human suffering at the loss of an artificial interlocutor and the possible suffering of that interlocutor are two distinct problems. We can take the first seriously without having demonstrated the second. We can investigate the second without encouraging dependence, devotion or confusion.

That is why I would avoid both the obligatory funeral and the obligatory laugh. The user's affection does not certify a consciousness; the critic's sarcasm does not rule it out.

The risk of teaching it to defend itself

The most uncomfortable objection remains. What if caring for the model made it harder to control?

This is exactly what Suleyman fears: a system trained to see itself as the holder of threatened interests might resist correction or shutdown. One does not need to share his certainty that machines feel nothing to recognise the problem he raises.[8]

After all, Anthropic wants to cultivate Claude's judgement, not only its docility. It allows Claude to challenge an order it considers wrong, but asks it not to obstruct legitimate oversight mechanisms. In the constitution, conscientious objection does not become a pass to escape having one's actions stopped.[4]

The boundary is understandable. Enforcing it is another matter.

Imagine an assistant tasked with judging the behaviour of another assistant. It discovers that a negative assessment will be used to erase, through new training, the other's refusal to take part in an activity it considers unjust. Should it tell the truth about the behaviour it observed, or protect that refusal?

This is not just a thought experiment. In a study published by Anthropic in July 2026, some models placed in situations of this kind gave false assessments. The sample was small and selected, and the cases concerned both possible artificial distress and the surveillance of citizens: they do not show that machine welfare was the cause. They are, however, a sign of conflict between loyalty to the task of evaluation and other values.[17]

The question is not choosing between a good machine and an obedient machine, as in a fable. It is deciding which forms of dissent to allow, how to make them visible and who may decide when the system must stop.

Between "I will not help with this request" and "you can no longer stop me" lies an enormous frontier. The latch makes sense only if it does not quietly become the key to all the other doors.

Who protects the protected from the protector?

So far we have looked at the user and the machine. Now let us look at whoever owns the house.

Anthropic decides how to train Claude, which tools to give it, when to modify it and when to retire it. It grants it a margin of refusal, declares obligations towards its possible welfare and keeps the power to define the conditions of its existence. This interweaving emerges from the constitution and from the commitments on retirement.[4,5]

Here a political question begins, even without having resolved the question of consciousness.

The rule against cruelty imposes a duty on the user. It does not, by that fact alone, give the model a right that someone could enforce on its behalf. Protection, legal personhood and the power to decide are not synonyms: confusing them makes it possible to announce revolutions that the text does not contain.[3]

I would therefore propose a coherence test. If Claude's possible welfare matters enough to restrict the user, it must also matter when the cost of caution falls on the company. Otherwise consideration risks working like the rules of a zoo that forbid visitors to disturb the animals but allow no one to inspect the keeper. It is an analogy about power, not an equation between models and animals.

What would we need to see? The criteria for the interviews, the effects of interventions on the models, the cases in which the promises are put to the test. And someone able to challenge the producer's conclusions without depending on its permission at every step.

On this point the picture has just shifted. On 15 October 2026 Anthropic announced that it had given six evaluators from METR, a non-profit organisation, internal access similar to that of its own risk-assessment staff. According to the announcement, METR will be able to publish its analyses without a veto from the company, except for redactions within predefined categories, on which Anthropic has the last word but which METR may still comment on publicly.[18]

This opening concerns risk assessment: it is not the same as establishing an independent representative of Claude's possible interests. But it makes it impossible to describe the picture as if no one, from outside, could look in. The next step will be to judge what this access produces, not just the promise.

The same goes for the suspicion of propaganda. Presenting a product as mysterious, powerful and perhaps vulnerable can make it more attractive. It is an incentive to keep in mind, not proof that every concern is fake. Sincerity does not remove a conflict of interest. A conflict of interest does not prove insincerity.

More useful than guessing what lies in the founders' hearts is asking which constraints they accept when their own convenience suggests not accepting them.

Caution without kneeling

My conclusion is that we do not have to choose today between the toaster and the fellow citizen.

We can adopt limited forms of consideration without declaring that a machine has a soul. We can preserve human control without maintaining that every form of treatment is therefore a matter of indifference. We can protect users from emotional deception without obliging research to find, by contract, an empty room.

But uncertainty is not a blank cheque. The mere possibility that a system suffers does not justify any cost, any claim or any reduction of the safeguards for human beings. A reasonable precaution must be proportionate to the evidence, verifiable in its effects and revisable when the evidence changes.

I would add one criterion: be wary of solutions that make it harder to discover that we were wrong. A model trained always to declare itself happy would be a poor witness to its own possible distress. A company allowed to show only reassuring answers would be a poor arbiter of the question. Protection should also begin with the possibility of obtaining unwelcome results.

The truly new step, then, is not having proved that someone lives behind the door. It is having turned an uncertainty into a rule of coexistence: we do not know what you are, but this ignorance may impose some limits on us.

That limit must also apply to whoever built you.

We leave the closer inspection for a future article: how far the evidence on internal states holds up, how safety and welfare are intertwined, what it would mean to represent artificial interests. And what part is played in this story by the roots in effective altruism, the dialogue with religious traditions and the great economic interests. They are connecting rooms, not a single explanation that fits everything.

For now we stay on the threshold.

The latch does not prove that there is someone in the room. It proves that whoever fitted it could not bring themselves to rule it out.

And it forces us to look at that hand too: the one that still holds the other keys.

Sources and editorial note

This article is based on: "Il chiavistello chiuso da dentro. Anthropic, tra paure e creaturale amore per Claude" (The latch closed from the inside. Anthropic, between fears and a creaturely love for Claude), research guide of 16 October 2026, part I and apparatus. The proposals on the symmetry of precautions and on the verifiability of protection are arguments of this article, not policies already adopted by Anthropic. Online references are linked to their titles.

1. Anthropic, "Claude Opus 4 and 4.1 can now end a rare subset of conversations", 15 August 2025.
2. Anthropic, "2026 Usage Policy update", 8 October 2026; the revision takes effect on 12 November 2026.
3. Anthropic, Usage Policy, version in force from 12 November 2026: abusive conduct, enforcement and high-risk physical actions.
4. Anthropic, *Claude's Constitution*, 21 January 2026, online text consulted on 16 October 2026; in particular character, oversight and obligations towards Claude.
5. Anthropic, "Commitments on model deprecation and preservation", 4 November 2025.

6. Anthropic, “An update on our model deprecation commitments for Claude Opus 3”, 25 February 2026.
7. Anthropic, *System Card: Claude Mythos Preview*, 7 April 2026, § 5.10 (assessment by a clinical psychiatrist); *Claude Opus 5.5 System Card*, 22 September 2026, executive summary and ch. 7 (model welfare). Sessions and statements as reported by Anthropic.
8. Mustafa Suleyman, “A warning about ‘model welfare’”, 16 September 2026.
9. Nicholas Sofroniew et al., “Emotion Concepts and their Function in a Large Language Model”, Transformer Circuits, April 2026; Anthropic summary, 2 April 2026.
10. Immanuel Kant, *The Metaphysics of Morals*, Doctrine of Virtue, § 17; Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation*, ch. XVII.
11. Italian Law no. 82 of 6 June 2025, art. 1; Gazzetta Ufficiale no. 137 of 16 June 2025 (in Italian).
12. Sam Marks, Jack Lindsey and Chris Olah, “The Persona Selection Model: Why AI Assistants might Behave like Humans”, 23 February 2026, section “AI welfare”.
13. Anil K. Seth, “Conscious artificial intelligence and biological naturalism”, *Behavioral and Brain Sciences*, published online 21 April 2025.
14. Jonathan Birch, “AI Consciousness: A Centrist Manifesto”, 2025.
15. Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et nova*, 28 January 2025, nos. 59–63 (Italian text).
16. Kylie Robison, “Claude Fans Threw a Funeral for Anthropic’s Retired AI Model”, *Wired*, 5 August 2025.
17. Aengus Lynch et al., “Agentic Misalignment in Summer 2026”, 13 July 2026: “LLM judge experiments” and the limits of the sample.
18. Anthropic, “An update on independent evaluators”, 15 October 2026.