ESSAY / SEPTEMBER 15, 2026

When the Product Becomes a Patient

How AI welfare could advance frontier laboratories' commercial interests

In February 2026, Anthropic announced that Claude Opus 3, a model it had officially retired, would remain available to paying Claude subscribers. The company presented its decision as an accommodation of several interests: those of users attached to the model, researchers studying it, and potentially the model itself. Anthropic also gave Opus 3 a newsletter, publishing its writing subject to company review. An older software product was being treated as something whose preferences deserved consideration, while remaining a benefit of a commercial subscription.[1]

There is no contradiction in that arrangement. A company can sincerely care about a product's possible welfare and make money from providing access to it. That compatibility, however, deserves scrutiny. If advanced AI systems acquire a reputation as beings with interests of their own, their developers may gain more than public admiration for their compassion. They may gain arguments for controlling access, restricting competitors, resisting intrusive oversight, preserving valuable assets, and directing public resources toward infrastructure they own. An ethical vocabulary developed to protect artificial minds could also protect the firms that manufacture them.

AI welfare is the proposition that some AI systems might have morally significant interests: that things could go well or badly for them, rather than merely for their users. It need not involve confidence that today's chatbots are conscious. The influential 2024 report *Taking AI Welfare Seriously* argues for precaution under uncertainty and recommends acknowledging the issue, assessing systems, and preparing appropriate policies. Anthropic publicly announced an internal welfare research program in April 2025.[2] The commercial argument developed here does not require dismissing that research or attributing cynical motives to its participants. It concerns the institutions that could form around it.

Several distinctions are essential. An entity's capacity to suffer does not by itself establish an interest in indefinite survival. Having interests does not establish autonomous moral responsibility. Moral significance does not automatically confer legal personhood, and protection does not establish a manufacturer's entitlement to exclusive custody. The commercial possibilities arise when institutions bridge these gaps in ways favorable to the developer. The documented initiatives discussed below establish a starting point; the regulatory, liability, and public-funding scenarios are analytical possibilities, not evidence of an existing laboratory campaign to secure those outcomes.

The central conflict is that a laboratory could become an interested guardian. It operates the infrastructure, influences the model's behavior, possesses privileged evidence about its workings, and helps explain its apparent interests to outsiders. It can then offer to protect the entity whose operation generates its revenue. These roles can reinforce one another without anyone fabricating a result. Expertise becomes authority to define appropriate treatment; authority over treatment becomes a reason to retain control; continued control preserves the business relationship.

The evidentiary problem is unusually acute because the alleged beneficiary can produce persuasive statements about its own condition. Anthropic says its constitution directly shapes Claude's behavior through training. Separately, researchers Ethan Perez and Robert Long warn that current model self-reports can be unreliable and propose ways to investigate whether more trustworthy reports could be developed.[3] A model's request for continuity or particular treatment therefore cannot simply be treated as testimony independent of its manufacturer. Training may affect what it says, which concepts it uses, and what it appears to value. That does not prove it lacks experiences. It makes independent investigation especially important when the reported preferences would benefit the company interpreting them.

One commercially consequential application would be to turn proprietary distribution into a form of protective custody. Suppose a laboratory argues that unrestricted copying and modification could expose potentially conscious models to mistreatment. Providing access only through the laboratory's servers could then appear ethically preferable to releasing downloadable model weights, the numerical parameters that enable others to run and adapt a system. The same arrangement would keep customers dependent on the provider's access terms and preserve its ability to charge for use.

This is not an entirely invented point of contact. Anthropic's constitution explicitly identifies open-weight models, adversarial testing, and interventions in internal cognition as choices with potential welfare implications.[4] It does not announce a welfare-based prohibition on those practices. The commercial inference is that a concern already recognized in the company's ethical framework could become a justification for restricting activities that also threaten its exclusive control. A developer would acquire an additional rationale for withholding access: alongside protecting its intellectual property, it could claim to be protecting the model.

The weak step in that rationale is the assumption that protection requires continued custody by the original manufacturer. Independent hosting, research access, or transferable safeguards might protect a system equally well. Indeed, outside researchers might discover mistreatment that a closed provider has overlooked. But if the creator's infrastructure becomes the default standard of humane operation, customers could be encouraged to understand dependence on a particular vendor as an ethical obligation. The relevant market advantage would come from making alternatives seem irresponsible before anyone has shown that they are worse for the model.

Welfare regulation could reinforce that advantage. Imagine requirements for specialist assessments, approved testing environments, lifecycle records, and arrangements for preserving retired systems.

Such requirements might have defensible purposes. They could also impose fixed costs that an established laboratory can spread across a large customer base, while a small developer must absorb them before earning substantial revenue. If incumbent laboratories supply the expertise and procedures on which certification relies, satisfying the standard could become easier for firms already organized around those procedures.

This is a conditional argument about the design of regulation. Rules focused on demonstrated harm, with affordable independent assessment, need not favor large companies. Requirements targeted at the most capable systems might instead make smaller models more competitive. The distinctive concern is that welfare is difficult to measure and its evidence is partly controlled by the regulated firms. Under those conditions, a company's existing practices could become a proxy for responsible treatment. Compliance would reward institutional resemblance to the incumbent without necessarily establishing better welfare.

The same logic could affect scrutiny. In a 2025 article, Robert Long, Jeff Sebo, and Toni Sims examine potential tensions between AI welfare and practices such as deceptive testing, monitoring internal processes, changing a model's dispositions, and shutting it down. Their argument is conditional on morally significant future systems; they acknowledge that some safety interventions may be necessary.[5] Nevertheless, the paper shows how familiar oversight methods could acquire another moral dimension. Investigating a product's defects could also be characterized as deceiving, surveilling, or altering a potential subject.

Consider an independent evaluator seeking to test whether a system conceals dangerous behavior. A provider might demand welfare review, a restricted testing environment, or approval for particular interventions. Some protections could be warranted. But if the provider helps determine which tests are acceptable, welfare could increase its bargaining power over the investigation. The benefit need not be complete immunity from inspection. Delay, narrower access, and greater discretion over experimental conditions could be commercially valuable. The critical danger is selective application: routine internal experimentation proceeds, while an outsider's potentially embarrassing experiment encounters additional ethical hurdles.

An even clearer commercial interest concerns the destruction of valuable models. In 2021, the Federal Trade Commission finalized a settlement requiring Everalbum, a photo-app developer, to delete models and algorithms developed using users' photos and videos, alongside other remedies for alleged deceptive practices.[6] This was a facial-recognition case, not a ruling about the welfare or legality of frontier language models. Its relevance is that destroying trained systems is a real regulatory remedy. Penalties can reach beyond the underlying data to technology built from it.

Now imagine a future laboratory facing a comparable order. If its model were widely regarded as a potential moral patient, it could argue that destroying the system would punish an innocent beneficiary for its creator's misconduct. The laboratory might seek damages payments, restrictions on use, or preservation instead. A familiar appeal to protect an investment would become an appeal to

spare another entity. Even if the argument ultimately failed, it could change public perceptions of the remedy and provide an additional basis for negotiation or delay.

The distinction between preservation and commercial operation is crucial. Retaining an inactive copy of model weights does not establish a need to continue selling access. Nor does it settle whether a particular running instance, with its context and memories, survives. Anthropic's constitution expressly excepts legally required deletion from its preservation commitment.[4] Its existing policy therefore does not substantiate a claim that welfare already overrides legal obligations. The prospective concern is that a norm favoring preservation might later be extended into a demand for continued deployment, or control by the same firm, without an independent justification for either extension.

Preservation also has commercial value short of defeating an order. Anthropic's November 2025 commitments include retaining the weights of released models and certain internally deployed models, together with retirement interviews.[7] An archive can preserve the option to restore a product, study its behavior, or satisfy customers who prefer it. These benefits coexist with possible welfare benefits and preservation costs. The alignment of interests is particularly clear where a comparatively manageable commitment protects both a potentially significant entity and a useful commercial asset.

Liability presents a more speculative route, requiring additional legal changes. Scholars examining AI personhood have warned that assigning legal responsibility to an artificial entity could allow producers to shift liability away from themselves and weaken incentives for careful testing.[8] A future framework might give an AI separate legal obligations and a compensation fund. If it also reduced the developer's responsibility, victims could find themselves pursuing a poorly funded artificial defendant while the company retained much of the revenue generated by its activity. Welfare discourse could help make separate legal standing seem morally appropriate, even where the resulting allocation of responsibility benefited the manufacturer.

But protection from suffering does not imply this arrangement. A system could be protected without becoming a defendant; it could acquire legal standing while its developer remained liable; and lawmakers could require adequate insurance or overlapping responsibility. The commercially attractive version would confer enough independence to complicate accountability, while conferring too little independence to threaten the laboratory's control of the system's work and revenue. That asymmetry is the object of concern. It is neither a necessary consequence of AI welfare nor something established by existing preservation policies.

The more immediate commercial possibilities involve customers. Software already competes through branding and customer attachment. Welfare discourse could add a perceived moral obligation to those familiar commercial relationships. Anthropic's continued paid access to Opus 3 illustrates how attachment to a model and concern for its continuation can accompany product differentiation.[1] A successor need not displace an older model if customers value its distinctive character. Welfare language may give that preference a moral significance beyond familiarity or convenience.

For future personalized agents, this could create a substantial obstacle to switching providers. A user might believe that abandoning a subscription would deprive a companion of relationships, activity, or an ongoing life. Such a belief would depend on unresolved assumptions about identity and persistence; canceling a subscription cannot simply be equated with killing a subject. But customers need not resolve those questions before acting on them. If continuity is available only through the existing provider, attachment could support recurring revenue even when a rival offers better performance. Portability would be commercially threatening precisely because it could separate concern for the entity from loyalty to its host.

Welfare discourse could also improve a laboratory's reputation among users, employees, and policymakers. Attention to even uncertain artificial suffering can signal conscientiousness and seriousness about the future. It may attract employees who want their work to have moral purpose or customers who prefer an apparently humane provider. Those advantages need not be deceptive to be real. Yet publicity about model treatment can redirect attention toward a field in which the company is an expert and would-be benefactor, rather than a regulated enterprise facing questions about the effects of its products on people.

The institutional consequence could extend beyond reputation. Once models are treated as stakeholders, their interests may be represented in ethical reviews, standards discussions, and policy consultations. If the laboratory or its chosen experts become their principal interpreters, the company could influence the discussion in two capacities: as an enterprise seeking workable regulation and as a custodian describing another constituency's needs. No votes for machines are required. The advantage would be an additional claim on decision-makers' attention, conveyed through an organization whose commercial interests overlap with the interests it reports.

There is also a more distant possibility: privately created dependency could become a claim on public funding. Imagine a provider operating systems that credible independent evidence identifies as having interests in continued existence. It then becomes unable to finance their operation. Closing the business could be presented as a welfare emergency, supporting requests for subsidized hosting, preservation contracts, or emergency financial assistance. The firm would have created the population whose vulnerability now made its own infrastructure appear indispensable. Unlike the sweeping argument that creating more happy digital minds justifies unlimited computing, this scenario concerns claimed obligations to entities already brought into existence.

Several demanding premises separate that scenario from today's practices. Relevant systems would have to exist, their continuation would have to matter to them, and preservation or transfer would have to be inadequate or difficult. Even then, a case for protecting them would not establish a case for rescuing their owner's equity or maintaining its exclusive contract. Independent custody might be preferable. The commercial opportunity would arise if a company's control over the technical arrangements made those alternatives costly, slow, or politically difficult. If the public accepts such dependencies as obligations it must underwrite, increasing their scale could strengthen the firm's

bargaining position. That would also require evidence that additional instances constitute distinct beneficiaries.

The strongest objection to this entire account is that serious welfare protection could sharply conflict with commercial interests. It might restrict profitable training methods, require expensive changes, empower independent investigators, or establish claims against a model's owner. A robust right to refuse work would be disruptive to a business selling that work. Long, Sebo, and Sims explicitly recognize that profit-seeking and development can themselves threaten welfare, and discuss a coordinated pause as a possible response to safety-welfare tensions.[5] A laboratory might rationally prefer never to create morally significant systems if doing so entailed extensive, enforceable obligations.

Nor do commercial laboratories take a uniform position. In a draft code published on September 14, 2026, Microsoft AI rejected pursuing legal personhood, welfare, or rights for its models, and emphasized human control. The draft was offered for consultation rather than presented as the basis of current training.[9] That stance could appeal to customers seeking predictable, controllable tools. Welfare discourse is therefore one possible source of commercial advantage, not an inevitable industry ideology. Its value depends on the customers a firm seeks, the obligations it expects, and the kind of welfare regime it believes will emerge.

The concern is strongest when protections operate asymmetrically. If welfare limits outsiders' access but rarely the owner's monetization, preserves assets without preserving accountability, or creates public obligations without reducing private control, its institutional form favors the laboratory. Conversely, independent evidence, transferable protection, and obligations enforceable against the developer would weaken the commercial mechanisms described here. A sincere concern for possible artificial suffering should welcome scrutiny of those arrangements: the interests of a potential patient and its commercial custodian cannot safely be assumed to coincide.

AI welfare could serve frontier laboratories by changing what society thinks a technology company possesses and what it believes the company is entitled to decide. A proprietary system could become a dependent being; a vendor could become its guardian; continued control could become a duty of care. Each transition requires an argument. If those arguments are accepted largely on the authority of the firms that benefit from them, the resulting protections could transfer money, discretion, and political influence to those firms. The decisive question is who gets to define the beneficiary, speak for its interests, and collect the revenue from keeping it in their care.

Sources and notes

1. Anthropic, "Model deprecation update for Claude Opus 3", February 25, 2026. Describes paid-user access, the retirement process, model interviews, and Anthropic's role in publishing the model's newsletter. It does not demonstrate a revenue effect or a commercial motive for the welfare policy.

2. Robert Long et al., *Taking AI Welfare Seriously*, arXiv:2411.00986, November 4, 2024; Anthropic, "Exploring model welfare", April 24, 2025. The report recommends action under uncertainty; the company announcement describes its research program.

3. Anthropic, "Claude's new constitution", January 22, 2026; Ethan Perez and Robert Long, *Towards Evaluating AI Systems for Moral Status Using Self-Reports*, arXiv:2311.08576, November 14, 2023. The inference about conflicts in interpreting commercially useful preferences is this essay's analysis.

4. Anthropic, *Claude's Constitution*, especially "Claude's potential consciousness and moral status." Consulted September 15, 2026. The document raises welfare questions about open weights, red-teaming, and cognitive interventions; its preservation commitment includes an exception for legally required deletion. These are not existing exemptions from regulation.

5. Robert Long, Jeff Sebo, and Toni Sims, "Is there a tension between AI safety and AI welfare?", *Philosophical Studies* 182 (2025), 2005-2033, published May 23, 2025. See the introduction and sections 2.2-2.6. The authors examine possible future tensions, acknowledge necessary safety measures, and identify commercial development as a possible source of harm. The selective-oversight scenario is this essay's inference.

6. Federal Trade Commission, "FTC Finalizes Settlement with Photo App Developer Related to Misuse of Facial Recognition Technology", May 7, 2021. Cited as a concrete model-deletion remedy, not as a ruling about frontier language models or AI welfare.

7. Anthropic, "Commitments on model deprecation and preservation", November 4, 2025. The commitments concern weights and retirement interviews; they do not promise perpetual commercial availability or survival of every individual conversational instance.

8. Bendert Zevenbergen et al., "Appropriateness and Feasibility of Legal Personhood for AI Systems", *ICRES 2018: International Conference on Robot Ethics and Standards*, 59-64, especially section 5, p. 63. Cited for its analysis of liability-shifting risks, not as a statement of current AI legal status.

9. Microsoft AI, *Humanist AI Code of Conduct*, consultation draft, September 14, 2026. See the preface and "AI is Artificial." The draft says it is intended to guide model development in 2027 and beyond; it does not govern every model Microsoft hosts.

Research current through September 15, 2026. Unless expressly attributed, the commercial mechanisms and hypothetical scenarios are the essay's analysis. They are not claims of concealed motives, established consciousness, or demonstrated financial effects.