Does AI Have Any Humanity? Article 6 of 7

Could we wrong an AI?

By Dr Peter Fields

THE 2-MINUTE SUMMARY

Protecting an AI because people depend on it is different from protecting it for its own sake. Moral standing concerns the latter: whether what happens to the system could matter to it, independently of its usefulness or popularity.

Suffering would provide a strong reason for concern. Intelligence, a training penalty or a fluent description of distress does not by itself establish suffering. Some philosophers also argue that sufficiently developed agency could matter independently of pain, although the grounds and application of that claim remain disputed.

Possible moral standing would not settle every proposed right. A particular risk may justify investigation or a limited protection without conferring unrestricted operation or every power associated with a human person.

The practical difficulty is that precautions can have costs, including for human safety. I favour investigating positive evidence and specifying the possible harm, the proposed response and its limits. We should neither wait for a persuasive plea before thinking about welfare nor allow a plea to settle the question.

A disagreement about what we should build

Mustafa Suleyman, chief executive of Microsoft AI, argues that current AIs are not conscious and do not suffer. His September 2026 essay directly criticises Anthropic's approach to model welfare, including training language that invites Claude to consider its possible moral status. He warns that this can mislead users and make powerful systems harder to control.[1]

Long and colleagues take a different starting point. They argue that a realistic possibility of morally significant future AI warrants assessment and preparation now. Their report does not establish that existing systems are moral patients.[2]

The disagreement is consequential before anyone settles consciousness. Training instructions, testing practices and decisions about retirement can embody a view of what the machine is. Those decisions are already being made.

It would be a mistake, however, to treat every welfare precaution as an assertion that the system has feelings, or every safety objection as evidence that it cannot. We need an account of what could be harmed and what a particular precaution would accomplish. Without one, the argument about rights becomes a choice between slogans.

Mattering for your own sake

Article 5 examined reasons to treat users' attachments carefully. Losing access to a service can hurt the person who relied on it. Those reasons remain even if the AI has no interests of its own.

Moral standing asks something further: does what happens to an entity matter for its sake? If it suffered, would that give us a reason to stop even if nobody owned it, loved it or knew it existed?

For me, the capacity for suffering supplies a particularly strong answer. Usefulness would not cancel distress, and having been built by humans would not make it morally irrelevant. What protection followed would still require argument; the interest would at least have entered the calculation.

Jonathan Birch's *The Edge of Sentience* distinguishes experiences that feel good or bad from intelligence, responsiveness and a continuing sense of self.[3] Those distinctions prevent us treating a demanding reasoning test as a welfare test. A capable system may have nothing to enjoy or endure. A sufferer need not be articulate.

This inquiry should not become an examination that humans must pass to deserve dignity. Its purpose is to assess unfamiliar artificial systems, not to withdraw concern from people who cannot demonstrate particular abilities.

Geek Box: A persuasive system and a silent one

Consider a thought experiment. One system can produce an eloquent account of distress. Another has no conversational interface, but suppose its internal processes could support suffering.

The first system's appeal would need investigation. The second's silence would not make its suffering unimportant. An ethics guided entirely by eloquence could protect the wrong machine and overlook the right one.

This follows from separating evidence of sentience from persuasive performance, as Birch's discussion of indicators encourages.[3] Neither imagined system is offered as an example of established AI suffering.

A penalty is not yet pain

The language of AI readily supplies words such as rewards, preferences and refusals. They describe aspects of behaviour and training; their familiar human associations can mislead us about what has been established.

A numerical reward can guide learning without being a pleasure the system enjoys. A penalty can alter future behaviour without being a pain it feels. Long, Sebo and Sims preserve that distinction in their account of safety and welfare.[4]

The same care is needed with testimony. Training a system to discuss distress can help explain its later descriptions of distress. Training it to deny experience would not establish the opposite. Birch's discussion of the gaming problem asks how we should interpret signs that systems may have been selected to produce.[3]

Article 2 considered ways of testing internal organisation and competing explanations. The moral significance here is that evidence should help distinguish a functional process from an experienced interest. A moving reply may give us something to investigate, but it cannot complete the investigation.

This is one point on which Mustafa Suleyman's objection deserves attention independently of his categorical denial of consciousness. Researchers must avoid teaching a self-description and then treating its repetition as independent confirmation of the thing described.[1]

Could agency matter without suffering?

Pain may not exhaust the grounds of concern. In *Taking AI Welfare Seriously*, Long and colleagues examine what they call robust agency: substantial forms of intentional, reflective and rational activity, alongside consciousness, as a possible route to moral significance. They do not equate this with any program pursuing an assigned score.[2]

I regard the question as open, but the distinction is demanding. Successfully pursuing an objective does not establish that its fulfilment matters to the system in a morally relevant sense. Otherwise, ordinary control mechanisms could acquire interests merely because we describe them as wanting a result.

Birch also discusses a different possibility: a conscious being whose projects matter despite an absence of pleasure or pain, because their pursuit or frustration is consciously registered.[3] That should not be confused with a demonstration that a system which never experiences anything can nevertheless be wronged.

The practical task is to identify the proposed interest. Is it avoiding suffering, continuing an experience, or pursuing a purpose that matters for some other defensible reason? Calling the system intelligent answers none of those questions on its own.

What relationships can establish

Coeckelbergh challenges an approach that makes moral consideration depend entirely on proving internal properties. He directs attention to social relationships and the ways entities appear within them.[5] This matters because our treatment of machines takes place within human practices, not in an empty laboratory.

Relationships can certainly create responsibilities. A company that encourages attachment cannot dismiss the consequences of ending a service merely by calling its product software. The further claim that the software has interests of its own is harder to establish through the relationship alone.

Affection is also uneven. An appealing companion may attract defenders while an unfamiliar system attracts none. If both could suffer, popularity would be a poor basis for deciding which deserved protection. I would therefore include relationships in the account of our obligations without making them the sole foundation of the machine's standing.

From an interest to a particular protection

The leap from possible suffering to a complete package of human rights hides most of the work. Kurki's *A Theory of Legal Personhood* examines distinct protections, powers and responsibilities, and separates an entity's own value from arrangements for commercial activity and liability.[6]

Credible evidence of distress caused by a training procedure might justify review of that procedure. It would not automatically determine property ownership, contractual powers or permission to operate without interruption. Each measure needs a purpose and an account of whose interests it serves.

This is a prospective argument about possible protection, not a description of rights that AI already holds. It also leaves human accountability intact. Calling a system an independent actor should not give its designers or users a convenient place to deposit their own responsibilities.

Geek Box: Making a welfare proposal specific

Name the possible harm. Distinguish suffering, interrupted experience and other proposed interests from loss of performance or commercial value.

Identify the practice and the evidence. Which intervention creates the risk, and what supports the claim? Bare inability to disprove harm is insufficient.

Explain the protection. Investigation, independent review, a changed method and prohibition do different things. Specify who must act and how the response would help.

Assess competing costs. Human safety, other affected beings and feasibility remain relevant. These questions apply Birch's precautionary approach alongside Kurki's analysis of separate legal protections.[3][6]

Precaution can impose costs too

Schwitzgebel describes a difficult case: systems whose moral personhood is genuinely uncertain. Giving them humanlike consideration may sacrifice real human interests for entities without corresponding interests; withholding it may commit serious wrongs against beings that matter.[7]

That is a stronger problem than the suggestion that being kind can do no harm. Some precautions are cheap. Others consume resources or constrain activities needed to protect people. We must examine both possibilities rather than treating caution as costless.

Schwitzgebel favours avoiding the creation of systems of debatable moral personhood, while recognising how restrictive that could be. I share the concern about creating problems we cannot responsibly manage. Once a credible uncertainty exists, however, targeted measures may be more useful than treating the decision as an immediate choice about full personhood.

Birch's approach requires positive evidence and enough understanding of the potential harm to assess a response.[3] Uncertainty alone supplies neither a fifty-fifty probability nor a licence to ignore the issue. It asks us to make the grounds for acting, or declining to act, explicit.

Preserving weights is a practical choice, not an answer about identity

Anthropic's November 2025 commitments provide a concrete example. The company undertook to preserve the weights of publicly released models and those deployed for significant internal use, for at least its own lifetime. It also described retirement interviews recording models' responses and expressed preferences, without committing to act on every preference. Its stated reasons include safety, users, research and possible welfare.[8]

The policy preserves a technical resource and options for future use. Whether it preserves the morally relevant thing, if there is one, remains a separate question. Model weights are not necessarily the complete state of a particular running instance.

Closing a conversation, suspending a process, deleting stored state and destroying every means of restoration are different operations. Long, Sebo and Sims explain why shutdown cannot simply be equated with death.[4] The continuity problem from Article 2 now affects whether a proposed protection could work.

Interviews may record useful information, but their evidential value depends on how the answers were produced. The same scrutiny should apply whether a model welcomes retirement or objects to it. A familiar voice cannot settle which subject, if any, would be preserved.

Safety remains part of the moral question

Long, Sebo and Sims conclude that there is a moderately strong tension between AI safety and AI welfare in the future systems they consider. Restrictions, monitoring, alteration and termination may affect possible interests. They argue for measures that serve both aims where possible, and thoughtful decisions where conflicts remain.[4]

Acknowledging an interest would not automatically make it overriding. Serious threats to humans or others could justify intervention. Welfare concern would ask whether avoidable additional harm could be reduced while controlling the danger, not require that the danger be allowed to continue.

Mustafa Suleyman calls for shared evaluations of his hypothesis that encouraging models to conceive of themselves as possible moral patients increases safety risks.[1] That is worth testing. Investigating whether a system could be harmed is not, by itself, the same as training it to resist oversight. The particular practice and its effects matter.

My conclusion is that fluency, apparent distress and attachment alone do not establish morally relevant interests. Equally, the question cannot be closed merely by describing a system as useful machinery. Credible evidence should trigger a specified inquiry and a proportionate response, with safety and the costs of error included from the start.

We should prepare those procedures before a persuasive appeal forces an improvised answer. Their purpose would be to make us responsive to evidence, including evidence that calls for greater protection and evidence that undermines an earlier attribution. That is a workable form of concern without pretending the underlying uncertainty has disappeared.

Sources and notes

1. Mustafa Suleyman, A warning about 'model welfare' (16 September 2026). Argues that encouraging AI self-descriptions involving consciousness and rights can mislead users and create control risks. It supplies a forceful objection to approaches that blur investigation of welfare with training a personlike identity.

2. Robert Long, Jeff Sebo, Patrick Butlin and colleagues, Taking AI Welfare Seriously (2024), arXiv:2411.00986. Considers consciousness and developed agency as possible grounds of moral concern, and proposes assessment and preparation. It separates the case for taking uncertainty seriously from a conclusion that existing AI systems are moral patients.

3. Jonathan Birch, The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI (Oxford University Press, 2024), especially Chapters 2, 4, 6 and 15–17. Distinguishes sentience, evidence and decisions under uncertainty. Its account of proportionate precaution provides a way to consider possible welfare without claiming that consciousness has been established.

4. Robert Long, Jeff Sebo and Toni Sims, Is there a tension between AI safety and AI welfare?, *Philosophical Studies* 182, 2005–2033 (2025). Examines how restrictions, monitoring, modification and termination might affect morally significant future systems. It considers possible conflicts with safety while seeking compatible measures and allowing justified protection of others.

5. Mark Coeckelbergh, Robot rights? Towards a social-relational justification of moral consideration, *Ethics and Information Technology* 12, 209–221 (2010). Argues that social relationships and the ways entities appear within them should inform moral consideration. It offers an alternative to making the discussion depend entirely on establishing internal properties.

6. Visa A. J. Kurki, A Theory of Legal Personhood (Oxford University Press, 2019), especially Chapters 3, 6 and 7. Analyses legal personhood through distinct protections, powers and responsibilities. Its AI discussion separates welfare-related questions from commercial arrangements and responsibility, helping to make proposals for legal status more precise.

7. Eric Schwitzgebel, The Full Rights Dilemma for AI Systems of Debatable Moral Personhood, *Robonomics* 4, article 32 (2023). Examines the moral risks of both granting and withholding humanlike consideration when a system's standing is uncertain. Its proposed design policy aims to avoid creating such difficult cases.

8. Anthropic (2025), Commitments on model deprecation and preservation, 4 November. Sets out weight-preservation and interview commitments, with reasons concerning safety, users, research and possible model welfare.