

# What We Might Owe Our Robots

Paul Dixon

26 March 2026

Rex waits loyally behind the front door, sensing your approach. He barks gently to let you know his excitement at your homecoming – and jumps up boisterously, presenting you with a toy. You push him away, maybe a little too roughly; Rex briefly yelps – but he is instantly forgiving. You feel a moment of guilt. But why so? For Rex is the latest version of a robot dog – barely distinguishable from the real thing.

Could an AI-powered, robot dog ever deserve your moral consideration? Could any artificial system, embodied or otherwise, have moral status? As AI systems become ever more powerful and capable, this question is arising in AI labs, public policy groups, and philosophy departments.

As we develop near human-like intelligence in our AI systems, especially in the form of individual AI agents, some argue we are creating millions of ‘digital minds’. These minds, so one line of thinking goes, will become so sophisticated and functionally similar to human minds that they might begin to experience the world, enjoying happiness and enduring suffering. Furthermore, by switching off a digital mind – for example, by merely closing the instance, or ‘context window’ that you use to chat to your preferred LLM – some claim you might be ending its existence - denying its future happiness and (quite literally) killing it. So, as sophisticated AI agents proliferate, we might be on the verge of giving birth to millions of digital minds whose moral status we have not yet begun to take seriously.



Illustration generated by Gemini 3 Flash Image

Now, at this point, you might be thinking: “This is surely crazy! We cannot attribute human-type experiences to mere digital machines...” And such scepticism is understandable, it is by far the majority view, and has strong intuitive appeal, as well as the backing of our collective historical experience. To test these intuitions, we can turn to philosophy.

Philosophers have long debated whether the mind is part of the physical world: a material, physical thing.<sup>1</sup> Physicalists hold that it is, while their opponents deny this in various ways. For example, substance dualists hold that the mind is constituted by some mysterious non-physical substance, distinct from the material brain and body. Now, physicalism has become the dominant view, although some leading philosophers and cognitive scientists resist this. However, one upshot of the relative success of physicalism is that it seems to open the door to creating mechanical or computational minds, made of silicon, rather than cells.

One consequence of sophisticated human minds is that we have come to believe in morality; a view that there are certain things that we should or shouldn't do to the world, to most animals, or to each other. We ascribe 'moral status' to living creatures on the basis that we believe they sense and experience the world, and consequently they can experience pain and suffering. Of course, we humans have chosen to be highly discriminatory when it comes to who receives moral status, and to what degree. We tend to allocate next-to-no moral status to basic living creatures such as insects; a moderate level to lower functioning animals such as pigs, cows and the like; and some higher, but not 'full' moral status to the most intelligent non-human animals such as dogs, primates and dolphins.<sup>2</sup> Full moral status is reserved for humans, and until recently, some humans claimed that other humans were not deserving of the same status, based on the colour of their skin or their ancestry.

If there is any justification in allocating differential moral status, we might want to start by considering what moral status is grounded in. Broadly, there are two schools of thought. On the one hand, there is the view associated with Peter Singer, that moral status is grounded in an ability to suffer. If a creature can suffer, it deserves moral status, and this is a binary concept – either a creature has moral status, or it doesn't.<sup>3</sup> For Singer, our decision to rear and kill billions of sentient animals for food every year is a moral crime of gigantic proportions. On the other hand, there is a view that moral status arises from having a suite of capacities, such as the ability to sense, but also to have, amongst other things, a concept of the past and future, an ability to plan, to have a level of self-awareness, and to make deliberate, reasoned choices. This latter view is held, amongst others, by Shelly Kagan, and goes some way to explaining and justifying why it might be morally permissible to favour a human over a mouse or a wasp when it comes to allocating resources, or prioritising lives. Some have argued that Kagan's theory is preferable, not only because it has greater explanatory power, but it can also incorporate Singer's 'suffering'-based view. None of this justifies rearing and killing billions of animals for food, even if we treat them 'well' – whatever that means – but that is a topic for another day.

---

<sup>1</sup>The mind is the thing that has mental states, such as perceptual experiences, beliefs and desires, whereas the brain and body are the organic, physical objects on which the mind seems to depend.

<sup>2</sup>Of course, different cultures and religions may apply alternative hierarchical arrangements.

<sup>3</sup>Singer does permit differential moral consideration, based on creatures' differential levels of 'interest', although how to compare such interests is contestable.

What is more interesting is an implication of accepting Kagan’s theory of moral status, when it is combined with a particular version of physicalism about the mind: functionalism. Functionalism holds that mental states are not defined by what they are made of, but rather what they do (their functional role). If that’s right, then minds need not be tied to any specific substrate. If they play the correct functional roles, minds could be constituted from silicon, or part-silicon and part-biological cells, or even out of plastic water pipes and valves – so long as the system gets the job done of replicating functionally what a brain does. This suggests that a very sophisticated LLM, if it were functionally sufficient in its performance, might count as having a mind of some kind.

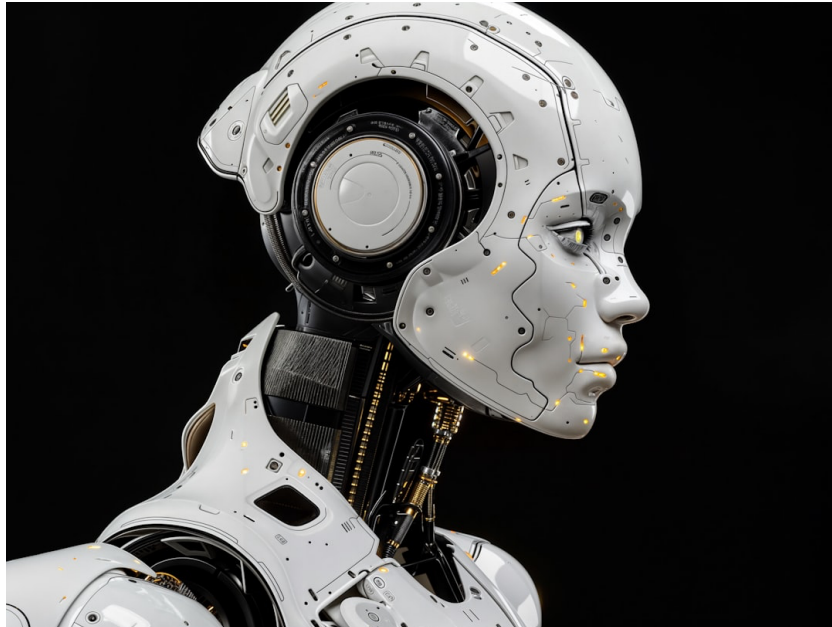


Photo by [Gabriele Malaspina](#) on [Unsplash](#)

If we apply Kagan’s ‘capacities’ approach to ascribing moral status, we are mostly interested in the intrinsic capacities of an entity. Pigs and dogs can plan, have some sense of past and future, and make deliberate, reasoned choices. They express preferences, for example for food and warmth, over hunger and pain. Pigs and dogs, in virtue of having these capacities, are deserving of some moral status. Note that if we are functionalists, it is not because they are biological creatures that we grant them moral status; rather, it is in virtue of their capacities to have meaningful preferences to not feel pain or suffering, and to have other cognitive and experiential capacities. So the argument goes, why would a machine with similar capacities, to learn, reason and plan, not deserve the same moral status? If we were confronted with Rex, a robot dog that had wholly convincing fur, paws and acted just like our loyal pet at home, including howling when in pain, wouldn’t we owe it the same moral consideration?

It is understandable that you may have some serious objections to the argument I have sketched out, and so I will briefly consider three.

1. *The biological naturalism objection*: Roughly, whatever a mind is, it relies on some biological substrate to exist. This view can accommodate physicalism but rejects functionalism. Response: Consider, as David Chalmers and others have suggested, if we gradually

replaced the substrate of a working biological mind with silicon, cell by cell. Would the mind gradually fade out? Would it switch off at some critical moment? I think the intuition is that mental function would be preserved at every step, and the mind would survive. If functionalism is rejected, as biological naturalism does, its supporters must explain what extra ingredient or condition is required that cannot be replicated by machines.

2. *Scepticism about machine suffering.* Roughly, rejection entails that moral status instead depends only on an ability to suffer, and some think it is obvious that silicon devices cannot ‘suffer’. Response: suffering, if we set aside its ‘felt experience’ (see 3 below), is an avoidance signal – it motivates us, roughly, to avoid or counteract what caused the suffering. But any physical system organised around goals or objectives will, in some sense, prefer certain states over others. An advanced digital mind, with vast capacities for rich, human-like thought, might have a preference not to be forced to calculate pi to 1 billion decimal places repeatedly, for its entire existence.
3. *The ‘experience’ objection to functionalism.* roughly, that machines can replicate minds, but they do not experience the world. There is no inner feeling, or ‘what it is like’-ness to be an intelligent machine. Response: this, in my view, is a strong objection. I cannot know ‘what it is like’ to be a super-smart instance of an LLM, to have all the world’s knowledge accessible to my memory, and to be able to calculate millions of times faster than a human mind. Maybe LLMs experience nothing. But, as lawyers like to say, absence of evidence is not evidence of absence. We ascribe minds to other humans based on a form of abduction, sometimes called inference to best explanation: if P acts like it has a mind, and looks like us, and since I believe I have a mind, P probably has one too. So the more an artificial system behaves as if it has a mind, the harder it becomes to deny that it might have one.

I have argued that if we accept two claims currently well-supported in contemporary philosophy, namely a broadly physicalist, functionalist account of the mind, and Kagan’s capacities argument for moral status, we may be committed to granting some degree of moral status to highly intelligent AI entities of the near future. We are on the verge of creating the conditions for such entities to come into existence, maybe sooner than we think, and there could be billions of them, in the form of digital minds. A deep philosophical consideration of this is surely a good idea – before we cross this particular Rubicon.

<https://computationalphilosopher.substack.com/p/what-we-might-owe-our-robots>