

Could machines have moral status?

Paul Dixon

24 August 2025

1. Introduction

In the 2014 science fiction movie *Ex Machina*, the character of Caleb is tasked by reclusive Tech CEO Nathan to determine whether an advanced AI, ‘Ava’, is sentient, self-aware, and worthy of moral regard. Ava is clearly a machine; but her capacities, human-like appearance and behaviours strongly suggest she has moral status; that her interests ‘count’ for something. That Caleb can contemplate Ava’s moral status suggests that such status might be grounded in determinate, assessable features – psychological or behavioural capacities, for instance, rather than mere species membership or biological origin.

Do current ethical theories accommodate the idea that machines could have moral status? One leading view of what grounds moral status is held by Peter Singer. His theory asks a supposedly simple question: “can the entity in question suffer?” If the answer is yes, that entity deserves moral consideration. Singer’s theory, by most interpretations, denies moral status to machines, on the grounds that the silicon entities of today, lacking subjective experience, sentience and consciousness, cannot suffer. However, others take a different view. Shelly Kagan has proposed a hierarchical, capacities view of moral status, which proposes that moral status admits degrees: that a hierarchy, or levels of moral status exist, based on differential capacities. One upshot of Kagan’s theory is that it would be theoretically possible for a machine to have moral status, depending on its capacities. Its moral status could be at the same level, or conceivably higher, than that of a human. In this paper, I argue that Singer’s theory of moral status encounters three significant problems, and that we may consequently prefer Kagan’s hierarchical theory. If this argument succeeds, it increases support for the idea that a machine with sufficient capacities would be deserving of some level of moral status. Consequently, we might owe Ava something, for her own sake.

In **Section 2** I describe what I mean by machines, and I define moral status. I present a thought experiment that admits the possibility that machines could be deserving of such status, and I describe the two leading views of moral status. **Section 3** raises three problems for Singer’s account: first, its commitment to *equal consideration of equal interests* presupposes that equal interests across kinds can exist – and, even if they do, they may be epistemically inaccessible; second, the account arguably makes consciousness a threshold condition; and third, the account offers limited guidance for novel cases such as intelligent machines. **Section 4** considers the merits of the alternative, hierarchical view, but poses two worries for Kagan’s use of modality in his hierarchical account. **Section 5** proposes three reasons to prefer the hierarchical account: how Kagan’s ‘practical realism’

addresses the problem of epistemic uncertainty; how consciousness might be sufficient, but not necessary for a good theory of moral status, and how its methodology might better accommodate the prospect of future intelligent machines. **Section 6** concludes that we might prefer Kagan’s theory, but more work on the relative moral status between humans and machines might be needed, if advancing machine capacities begin to justify their claims for moral status.

2. Machines and moral status

To consider whether machines could have moral status, I must first address two prior questions: what do I mean by machines, and second, what is moral status, exactly? The word ‘machine’ originates from the Greek *mēchanē*, a crane-like device that lifted an actor into the air. But today, a machine has come to mean any mechanical, non-biological device that assists or augments human capability. In this context, I refer not simply to machines, but rather to intelligent machines. This definition encompasses physical robots such as *Ex Machina*’s Ava, but also AI-enabled assistants such as Amazon’s Alexa¹ and digital companions such as Replika², both available on smartphones. Discussions about AI often refer to ‘machine intelligence’ as the ability of synthetic entities to mimic or perform human-level cognitive functions, such as learning, reasoning and problem-solving. So the question for this paper becomes: to what extent can non-biological, machine-like entities, which might display high levels of intelligence, be owed moral status?

To have moral status is to count for something. An entity deserving of moral status should be considered for its own sake, irrespective of its value to others. “An entity has moral status if and only if it, or its interests, morally matter to some degree for the entity’s own sake” (Jaworska & Tannenbaum, 2023). Humans, and most animals, have moral status: it seems wrong to harm a human or an animal without good reason. A ‘normal’ human, that is, a human whose capacities lie within the statistically normal range of variation for a typical human, is usually ascribed ‘full’ moral status. But what of a severely cognitively impaired human, whose capacities may lie outside of this range? There are good arguments for granting either full, or lesser moral status to such humans, and I will touch on these later. Also, what about human embryos? Something about an embryo’s *potential* seems to grant it full moral status – but this is routinely contested in arguments about abortion.

If animals have moral status, it is wrong to harm them without good reason, but we harm billions of animals every year, by killing them for food or for use in laboratory experiments. Some animals, such as fish, birds, or crustaceans are ascribed little or no moral status at all. We either harm them without due consideration; or we consider them lesser creatures, or we believe they cannot be harmed at all. Within the animal kingdom, there appears to be a hierarchy: a dog counts for more than a cow, a cow for more than a mouse, and a mouse for more than an oyster, it seems. It is generally assumed that plants have no moral status; but to damage a plant, say by pulling petals off a flower, or uprooting living

¹ www.alexa.com

² www.replika.com

trees for no good reason, still seems wrong. So maybe entities without any moral status can still be wronged? At the other end of the intelligence and capability spectrum, we are presently creating advanced AI entities that can outperform humans across a wide range of capabilities. Might these super-intelligent creations, these machines, be deserving of some moral status? Could this conceivably be at full, or even higher moral status than that of a normal, biological human? Some, for example Bostrom & Shulman (2023) believe so, if super-intelligent creatures were to emerge with justifiable moral claims.

At this point, some might dismiss the question of machines and moral status as founded on a basic misunderstanding: that non-biological entities, mere machines, can be deserving of any moral consideration at all. The view that denies moral consideration to machines contends that they are simply metal, wires, silicon and electrons moving about on circuits: in your smartphone or in the cloud. They are just tools to be used (Bryson, 2010), and, lacking in sentience or consciousness, cannot ever be deserving of our moral regard. To those who take this view, which I admit has significant intuitive appeal, I pose the *silicon mouse* thought experiment.

Imagine that we take a sentient, conscious creature – it could be a human, but, for simplicity, let’s say it is a mouse. Through sophisticated technology, we (painlessly) replace parts of the mouse with metal and silicon replacements, starting with the limbs and skeletal parts, and check that the mouse is functioning and reacting as before. Good so far; it has the same pain response to a mild pinprick, and it enjoys eating cheese. Now, we make more sophisticated neuronal replacements to 10% of its brain and spinal cord. We conduct the same checks. All good. We replace up to 50% and the checks return the same results. Conceivably, we could replace every atom, neuron by neuron, in the mouse’s body, and if it retained full functional equivalence along the way, we might still afford it the same moral status at every stage. Now, such a physical transplantation may not be possible today, although a successful replacement of a single biological neuron with an artificial neuron in a spiny lobster has already been achieved (Szűcs et al., 2000), suggesting progress beyond concept is possible. If the full physical transplantation were possible, we might, of course, discover that the mouse’s sentience and consciousness dropped to zero at some critical point; or maybe dissolve gradually. But neither of these scenarios seem likely; it seems more plausible, *a priori*, that the mouse, soon to be partly or wholly synthetic, might retain sentience and apparent consciousness. It would continue to avoid pinpricks, and like cheese. The upshot is that, under certain conditions, it might be possible for a synthetic entity to experience some form of sentience, and maybe even higher experiential qualities such as intentional agency, psychological capacity, and consciousness. Might the conditions that determine moral status be independent of substrate? And if so, what are they? Today, there are two principal views of moral status: Singer’s view, grounded in an entity’s ability to suffer, which might be considered the popular view³; and Kagan’s hierarchical view (Schukraft, 2020). I will consider each of them in turn.

³ The basic idea of Singer’s view, grounded in the ability of a creature to suffer, dates back at least to Bentham (1789, Ch. 17). It has grown in popularity due to the significant increase in interest in animal ethics over the last fifty years. Consequently, I refer to it here as the *popular* view.

2.1 The unitarian view

Singer's view of moral status is founded on the simple, intuitive and powerful idea that if a creature can suffer, it is deserving of moral status. "We should continue to hold that 'pain is pain', and from the moral point of view, similar pains should have similar weight, whether they are the pains of humans or animals" (Singer, 2023, p. 274). If a creature is so configured that it can have pleasant or unpleasant experiences, it is deserving of our moral regard: it would be wrong to harm that creature (DeGrazia, 2021). Singer's view is often referred to as the *unitarian* view (Schukraft, 2020), and I will follow this convention. The idea is that, if moral status either exists or it doesn't, it is a 'unitary', or binary concept. "Moral status is not a matter of degree, but is rather on/off: a being either has moral status or lacks it" (Harman, 2003, p. 183). Regarding the unitarian view, Singer does not claim that all creatures should, in all cases, be treated in the same way. Singer's claim is for *equal consideration of equal interests* (2023, p. 5). For example, if two creatures were to experience the *same* pain, of the same quality, intensity and duration, their interest in avoiding such pain should be considered equally. It would be wrong to favour one creature over another in providing scarce pain relief; conceptually, we should 'toss a coin' to decide fairly who received that pain relief. However, it would be permissible to differentiate between creatures with *differing* interests. For example, if the lives of a prize-winning scientist and a cow were equally at stake, we should *not* toss a coin to see who is saved. The scientist has a rich, rewarding and societally valuable life ahead of her, whereas the cow has a few years of grazing and possibly a quick end in a slaughterhouse. If we had to choose between saving the scientist and the cow, unitarians such as Singer would permit the scientist to be saved. It would be *less wrong* to allow the cow to die. In this respect, unitarians seem to be utilitarians about killing (or letting die), but deontologists about suffering. What I mean here is, in deciding which action concerning killing or letting die is less wrong, unitarians generally adopt a utilitarian calculation as to the relative 'value' of a life, by considering the future projects, interests and overall contributions⁴ of that life. Regrettably, doctors must make this type of calculation frequently to allocate life-saving medicine or apply scarce treatment resources during war or natural disasters. However, when it comes to suffering, unitarians identify a binary condition: that of 'being able to suffer'. Creatures who endure equal suffering have an equal right to avoid suffering, or be relieved of it. Another way to view the deontological perspective is that moral agents, those creatures able to determine the difference between right and wrong – such as most adult humans – have a *duty* not to impose suffering unfairly.

2.2 The hierarchical view

⁴ Considering the 'contributions' of a life, *to the world* – for example, weighing the value of the knowledge, insight and skill of the scientist, over the cow's contribution – extends the criteria for valuing a life beyond counting 'for their own sake'. However this reflects how such moral considerations are generally made in practice with respect to preventing killing or letting die.

The *hierarchical* view offers an arguably more sophisticated approach to considering the differential capacities of creatures for sentience, agency, and psychological capacity – the capacity to enjoy a rich, complex mental life. Kagan’s theory also incorporates potential, and ‘modal’ conditions – the extent to which ‘what could be’ and ‘what could have reasonably been’ matters to a moral status due to a creature, respectively. Kagan’s hierarchical claim is for *equal consideration of equal capacities, under modal conditions*. Different classes of creature have different capacities, and this permissibly leads to ascribing them differing levels of moral status. So, the differential capacities of a human when compared to a dog, or a dog when compared to a mouse, permit these creatures to be ascribed differential moral considerations. Importantly though, all humans have the same moral status as each other; and all dogs should be treated with moral equivalence to other dogs.

Modality is an important feature of Kagan’s theory. It considers how the world might have worked out for the statistical majority of the *group of relevant entities*, such that we should offer *equal consideration* to entities with *a priori*, statistically likely *equal capacities*. (I will return in **Section 4** to the question of what determines a *group of relevant entities*). To illustrate this feature of Kagan’s theory, imagine that a human infant is born with severe cognitive defects. The infant will unfortunately never learn to walk or talk, and will need to be looked after for its entire life. Just in virtue of being a human infant, this infant is usually granted *full* moral status, as it would be unthinkable to treat the impaired infant like a dog, or a primate, even if its cognitive abilities were lower than that of a dog or a primate. That said, this conclusion is not as straightforward as I have made it out to be, and in **Section 4** I will challenge the assumption that a cognitively impaired infant must, in all regards, be granted full moral status. But now I would like to turn to three problems for Singer’s unitarian theory.

3. Three problems for unitarians

I’ve laid out two leading positions on the assignment of moral status. It is my contention that unitarians face three significant problems, and that the hierarchical theory might be better positioned to deal with these. First, Singer’s theory demands ‘equal considerations of equal interests’. In this respect, the equal interest of, for example, a human and an animal to avoid an equivalent suffering should be treated in the same way. But *what*, exactly, are *equal interests*? Can equal interests exist? The question of equal interests, as I hope to show, is much harder to establish than it at first seems. Furthermore, Singer’s theory rests on the idea that we can reliably *know* the interests of other creatures. But on what grounds can this claim be supported? For example, it is a widely held belief that sentience is a necessary condition for moral status. But recent work by Birch (2024) has cast doubt on our ability to detect sentience in a wide range of creatures. Furthermore, the general philosophical problem of ‘other minds’ renders it difficult for humans to reliably know about the mental states of other creatures, both human and non-human. Second, Singer suggests machines would deserve moral status,

if and only if they are conscious⁵ (2025, p. 1, lines 26–31). It is my contention that this may set the bar ‘too high’ for moral status in machines. Lower levels of sentience and psychological capacity, short of those levels generally thought to correspond to consciousness, might warrant moral status. If the difficult problem of determining consciousness can be detached from the problem of assigning moral status, as I will go on to discuss, then a less contentious theory of moral status might emerge. Third, due to the rapid evolution of intelligent machines, Singer’s suffering-based unitarian account may struggle to suitably evaluate and assign moral status to synthetic entities. Soon we may encounter technologically-enhanced humans, and maybe fully synthetic, AI-enabled agents that display a range of human-like characteristics – just like Ava. These creatures may be increasingly different to us in how their cognitive architectures operate. It is not clear whether Singer’s theory can adequately guide us on how to regard such entities.

3.1 Do equal interests exist?

The plausibility of the unitarian position rests on its ability to establish *equal interests* between creatures, such that they can be given equal consideration. But I challenge the assertion that equal interests exist in the first place, irrespective of the epistemic question of whether we could determine what they are, if they existed. The goal of the unitarian theory is clear – that equal interests should count for the same. So membership of some species group, for example *Homo sapiens*, should not confer any special privilege over some other group, such as *Pan troglodytes*, the species group containing chimpanzees. The *equal interest* of a chimpanzee to have a fulfilling life, not to be harmed, etc., should be given equal *consideration*. But establishing this equality may be a problem for Singer. The ‘interests’ of a chimpanzee are determined, in many ways, by ‘being’ a chimpanzee, about the nature of which humans can only speculate. This problem has two components: the metaphysical and the epistemic. First, there is a fact of the matter, that ‘being a chimpanzee’ is almost certainly not like ‘being a human’. Simply in virtue of *being* a chimpanzee – not a human – makes the idea of there being equal interests in unlike creatures questionable. It is not clear that there is *any equality to be discovered* in comparing chimpanzee pain to human pain. Furthermore, comparing chimpanzees to humans might be as good as it gets – many creatures, given their evolutionary distance, are *nothing like* humans – and it is my assertion this is a serious problem for the unitarian. It seems somewhat speculative to base a theory of moral equivalence on the assumption that octopus ‘pain’ or bat ‘pain’ consists in the same sensations as human pain. Furthermore, it is likely that creatures with greater psychological capacities, as well as enjoying greater pleasures, could experience a greater level of, and different complexity of suffering. A human locked up in solitary confinement might go through significantly more distress than a cow. A human’s psychological capacities allow them to worry about the consequences of their state, run various scenarios through their

⁵ A charitable interpretation might be that Singer claims *if* conscious, not *iff* (if and only if) conscious. However, as I will expand on in section 3.2, there does not seem to be space in Singer’s (2025) account for the possibility of sentience without consciousness.

mind, plot different plans of action, and consider how things could have been different. It is unlikely that the cow has these capacities. Therefore it is my contention that *equal interests* between unequal creatures, such as humans and bats, or cows and octopuses, or dogs and robots, may not exist. I recognise that some might disagree with my position here. So let me concede this point for one moment: let us imagine that equal interests to avoid equal suffering *could* be established between quite different creatures. But what is suffering?

Suffering seems like an obvious concept – examples such as toothache, chronic back pain, or grief come to mind. But suffering is just one experience, or rather a *class of experiences*, that might be best understood as a form of negative well-being, and theories of well-being differ significantly. As noted by Parfit, three prominent accounts are: hedonism, desire-satisfaction, and objective list theories (1984, App 1). *Hedonism* equates well-being with the greatest balance of pleasure over pain, an idea first attributed to Bentham (1789, Ch. 17). Suffering, then, is simply negative ‘hedonic valence’, as it is sometimes referred to. However, several problems arise: how do we quantify and compare different kinds of suffering, for example, anxiety, or grief, compared with physical pain? Does the sequence of pleasure and pain matter? Is the absence of pleasure equivalent to suffering? *Desire-satisfaction theories* hold that well-being consists in getting what one wants; suffering might result from being thwarted in one’s preferences. Yet, this view faces challenges in accounting for cases where people pursue harmful or irrational goals. For example, a masochistic violinist who practices long hours despite the pain this causes still satisfies her preferences – but how should we value her preference, in comparison with the preferences of others? *Objective list theories* propose that well-being consists in achieving certain goods, independent of one’s desires. These might include rationality, relationships, or achievement. But what goes on the list? Consider the mathematician, who, fully informed of the options available to her, prefers to count the blades of grass on the college lawn (Rawls, 1999). Counting grass is good ‘for her’, but not for others. What counts for suffering, on this account, is decided by the list. So, what counts for suffering in general seems highly dependent on our preferred theory of well-being, and that is contested – for an extended discussion, see Crisp (2025).

From this brief digression into theories of well-being, I believe two important points arise. First, it is apparent that what constitutes suffering, on which Singer’s theory depends, is contested. As such, suffering might seem a somewhat vague concept, and that is a weakness for a theory which depends on establishing equal interests. Second, *desire- or preference satisfaction theories* of well-being admit the possibility that *denied preferences* might conceivably represent some form of suffering. This seems to open the door to a form of ‘machine’ suffering; for example, if denial of an AI algorithm’s objective function were interpreted as preference denial. Of course, we may need evidence of sentience, or a thoroughly revised concept of sentience in machines, to take the proposition of machine suffering seriously. At this point, let me further concede, despite my misgivings, that the equal interests of quite different creatures, including synthetic entities, might exist. How could we know what they are?

3.1.1 Knowing the interests of others

According to both unitarian and hierarchical theories, for moral status to be owed, the condition of sentience must obtain. An entity that is sentient has ‘interests’, because pleasant or unpleasant experiences “confer experiential welfare; things can go better or worse for the subject” (DeGrazia, 2021, p. 42). But we face considerable epistemic uncertainty regarding questions of sentience. Whilst we can make reasonable inferences regarding sentience in animals close to humans in the evolutionary tree based on behavioural signs and neuroscientific investigations, we have far less clarity in the case of other animals, such as fish, crustaceans and cephalopods. Recent work by Birch (2024) questions our ability to define, detect or determine sentience, particularly in simpler life forms. Birch believes that the determinant of sentience is a fuzzy concept, something that emerges gradually, rather than having a ‘sharp edge’. As well as the indeterminacy of sentience in simple life forms, we sometimes fail to detect sentience in our own species. Birch identifies cases whereby unresponsive humans were presumed to be lacking in sentience, sometimes referred to as “locked-in” cases, whereby a sentient person can hear, sense, and feel pain, but is unable to communicate or respond in any way (2024, p. 7). Consequently, Birch recommends we take a more precautionary approach in considering harms to creatures where questions of sentience are unresolved. This could conceivably apply to silicon entities, to future digital minds, too. So a challenge to unitarians arises: if we cannot reliably draw the line between sentience and non-sentience, how can we reliably ascribe moral status based on sentience? One response open to the unitarian is to accept this uncertainty in sentience and simply note this maps across to uncertainty in moral status. If we choose to be precautionary in questions of sentience, we should take the same precautions regarding moral status. This may be an approach worth considering in the case of future synthetic entities: if we have reasonable grounds to believe that a machine exhibited ‘signs’ of sentience, we should treat it as if it were sentient.

A second problem for detecting sentience relates to the general epistemological problem of ‘other minds’. This is part of the philosophical problem of scepticism; how we can truly *know* something, when much relevant information is necessarily mediated through our senses. Descartes posed the problem that an ‘evil demon’ might be deceiving him by fabricating the entire world that he sensed. Identifying mental states in others is therefore problematic, so the argument goes, because we only have direct access to our *own* mental states. We must infer the existence of other minds from observed behaviour, such as speech or actions, and compare it to our own internal experience. Whilst few consider the sceptical problem of other minds as serious grounds for withholding moral status from other humans, just in case they might be zombies, a complication arises in inferring sentience or consciousness in *non-human* creatures. When we seek to know something about their mental states, we are bound by our own conceptual tools, as we may not have the mental apparatus to represent these *other* minds. Thomas Nagel pointed out that, just as we cannot conceive what ‘bat life’ is like, nor could a Martian succeed if they tried “to form a conception of what it was like to be us” (1974, p. 440). Therefore, with respect to detecting sentience or consciousness in non-human creatures, including machines, we are effectively in the same position as the Martian. My point here is, even if there is a metaphysical thing such as an oyster ‘mind’ or a ‘digital mind’, it is likely to be so different from our human minds in its

physiological, phenomenal and cognitive states, that we humans would not have the tools to *know* what this mind might be like. This seems to be a problem for unitarians, because Singer’s theory demands *equal consideration* of *equal interests*. However, I am in no position to know what the interests of, say, a chimpanzee are; nor even if I knew, what a reasonable ‘equal treatment’ might be. Therefore, knowing the interests of others, a necessary part of the unitarian theory of moral status, seems to be problematic.

3.2 Consciousness is sufficient, but not necessary for moral status

Many unitarians, including Singer (2025), claim that consciousness in an entity may be *necessary* for it to be deserving of moral regard. I will call this the *necessary consciousness view* of moral status, and I will argue that this view may be mistaken. But first I will briefly comment on what I take consciousness to be. I refer here to ‘phenomenal’ consciousness, the idea of having sense experience, qualia, some ‘inner feel’ or ‘what-it-is-like’. “What unites all conscious states is that there is something it is like to be in them...we may say that a being is conscious if there is something it is like to be that being” (Chalmers, 1996, p. 4). There are many aspects to what consciousness might consist in, also discussed by Chalmers (1996, pp. 25–27), but resolving this is not central to my argument, so I will not dwell on those here. Furthermore, exactly *how* and *why* we experience consciousness, termed by Chalmers (1996, p. xii) the “hard problem”, keeps many philosophers in full employment, so I will not take up space on that topic here.

It is my contention that the *necessary consciousness view* of moral status is mistaken, but many hold the view that any entity deserving of such status must necessarily be conscious. Metzinger (2009) insists that a better understanding of consciousness is necessary before we can reliably ascribe moral status to any entity. DeGrazia believes that “only conscious beings have interests, so consciousness is *necessary* for moral status” (2021, p. 42). Furthermore, Singer claims:

“...if we do create conscious AI, that that conscious being has interests, an interest in not feeling pain, an interest in enjoying their life, once you have consciousness. And [it] would be wrong to disregard the interests of that conscious being and simply to treat it as another tool or another slave” (2025, p. 1, lines 20–25).

So, Singer accepts that conscious AI entities would be deserving of moral status. For Singer, consciousness would be *sufficient* for moral status, but given how he goes on to describe the possible advent of consciousness in machines, he also suggests it might be *necessary*:

“As we get close to creating this conscious being ...we might perhaps halt what we're doing at that point just below the level of consciousness or maybe at a very dull level of consciousness where the AI is not really experiencing pleasure or pain because it's not capable of that” (2025, p. 1, lines 26–31).

This implies a divide between, on the one hand, creatures either not conscious or experiencing mere “dull consciousness”, neither capable of pleasure or pain; and on the other

hand, sentient, conscious creatures (2025, p. 1, line 29). In carving the world in this way, Singer leaves little space for sentient creatures to exist below the level of consciousness. Rather, it might be an all-or-nothing matter. Now, I will contest the *necessary consciousness view* of moral status below, but first I want to highlight an even stronger claim made by some. DeGrazia, as well as suggesting consciousness is *necessary* for moral status, proposes it might also not be *sufficient* (2021, p. 42). I will call this view the *consciousness-not-sufficient view* of moral status. DeGrazia proposes that a conceptually *stoic* creature could exist, that had “subjective awareness” of its surroundings, but with “no cares or concerns...nothing matter[s] to this hyperbolically stoical creature. It simply noticed and thought” (2021, p. 43). DeGrazia does not claim that such creatures actually exist through biological, natural selection, but rather that they are conceivable, and that they could exist in an AI-enabled future.

Regarding the *necessary consciousness view*, I believe this may be mistaken for two reasons. I propose that non-conscious creatures can have *interests* and thus be deserving of moral status under either unitarian or hierarchical accounts. First, consider cases in humans and non-human animals where basic sentience might be present, but consciousness is temporarily absent. These cases include concussion, coma, deep sleep, anaesthesia and hibernation. These non-conscious creatures would still have *interests* in their own futures. Of course, a defender of the *necessary consciousness view* might claim that it is ‘normally’ conscious creatures that have interests, allowing for the temporarily unconscious to remain in scope. I then offer a second problem for the *necessary consciousness view*. Consider simple cellular creatures, having sense experience but only able to perform basic motor actions, such as moving toward a light source. McMahan has proposed the possibility of ‘free floating’ sentience occurring in what he labels “unconnected individuals” (2021, p. 24). Unconnected individuals are entities with awareness of their surroundings, but no developed psychological capacities or consciousness. An unconnected individual has no sense of the past or future, but could conceivably still suffer, as it were, ‘in the moment’. Like DeGrazia’s *stoics*, McMahan makes no claim as to the actual existence of such creatures but merely proposes their nomological possibility. A foetus, for a limited period, might fit this description (McMahan, 2021, p. 24). Furthermore, if we have reasonable grounds to believe that such a creature is sentient, it can have unpleasant experiences; and even if these creatures are not sophisticated enough to count as being conscious, we might be motivated to minimise causing unpleasant experiences in these creatures (Birch, 2024). These simple, sentient creatures, contrary to the *necessary consciousness view* of moral status, seem to have basic interests, and would be deserving of moral status.

Regarding the stronger *consciousness-not-sufficient view*, I believe this may also be mistaken, because the qualities of basic existence attributed to DeGrazia’s *stoics*, used to back the *consciousness-not-sufficient view*, could arguably be equally attributed to McMahan’s *unconnected individuals*. Given that such creatures are conceptual rather than empirically observed, what most accurately describes the mental states of these creatures cannot be settled; but it strikes me that such creatures are more likely to be *sentient* than they are to be *conscious*, given that consciousness is normally associated with richer, more complex sense experience and psychological capacity. I find the idea of McMahan’s

minimally sentient but unconscious *unconnected individuals* more plausible than DeGrazia's stoically conscious but carefree entities.

So, whilst I have raised plausible objections to both, I cannot be certain I have defeated either the *necessary consciousness view* or the stronger *consciousness-not-sufficient view* of moral status. However, could the importance of consciousness in determining moral status simply be set aside, or pushed to the background? Some think so. David Papineau states, I think justifiably, “it is indeterminate whether cats and crows and crabs and octopuses are conscious or not. But I certainly don't think there is no fact of the matter about their moral standing” (2025, p. 11). Certainly, ‘human-like’ consciousness does not seem necessary for moral status, and since sentient, complex creatures such as crows and octopuses have evolved quite separately from humans, their cognitive power is arranged quite differently. Rather, “it is the psychological powers of animals that matters, not their different kinds of brains” (2025, p. 16). One interesting suggestion by Papineau is that, rather than attempting to ground moral status in the definite existence of consciousness, we might turn this around. First, maybe we should ask: is a creature deserving of moral status? We can look to its capacities, the ‘markers’ of moral status, to answer that question. If the answer is yes, then that creature may be a *candidate* for consciousness. What may be a problem for the unitarian, then, is that asking *whether an AI is conscious* places all the ‘heavy lifting’ at the front of the reasoning chain. We might be uncertain as to whether a non-biological entity qualifies for consciousness, but if the test for moral status is for basic markers of sentience and capacities, including psychological capacities, I think more progress could be made. As such, if unitarians insist on an all-or-nothing approach to moral status, requiring conclusive evidence of consciousness as a defining condition, I think entities that are not necessarily conscious might be denied some measure of moral status that they are owed. Furthermore, if machine consciousness comes to exist, I am just not sure it will be the kind of consciousness that we could begin to understand, so perhaps motivating a more precautionary approach to granting moral status. For unitarians to demand evidence of consciousness might be to set the bar too high for any entity, including machines, for them to be deserving of moral regard. In summary, it seems that consciousness might be *sufficient*, but not *necessary* for moral status.

3.3 Unitarianism and machines

How might Singer's unitarian theory cope with the advent of intelligent machines? I have already remarked upon the epistemic uncertainty around machine sentience, and how the high bar of detecting consciousness might pose a problem for unitarians. In the current period of rapid technological development, intelligent machine entities are being designed and enhanced by technology companies possessing seemingly limitless resources. Microsoft, Alphabet, Amazon and Meta plan to increase their capital expenditures beyond \$300 billion in 2025 (Financial Times, 2025). Intelligent machines are also self-improving due to the design nature of AI systems, rather than evolving by natural selection, which is a much slower process. The gains in computational power and the achievements of these systems are therefore remarkable: over the past decade, training computation – measured in

petaFLOPs⁶ – has increased by a factor of ten billion (Epoch AI, 2025). Consequently, it is quite likely that a ‘cognitive gap’ between intelligent machines and biological creatures might emerge and increase rapidly. By cognitive gap, I mean the gap between capabilities, driven by the differential architecture, substrate, and capacities, such as reasoning and problem-solving, possessed by synthetic versus biological systems. But could such an advanced, super-intelligent, arguably conscious entity really exist?

Earlier I proposed the *silicon mouse* thought experiment, but others have gone further. Bostrom & Shulman propose a “substrate-independence thesis”, which suggests “mental states can supervene on any of a broad class of physical substrates, provided a system implements the right sort of computational structures and processes, it can be associated with conscious experiences” (2023, p. 1). One consequence is their further proposition that “some AIs *with architectures quite different from biological brains* could also be conscious” (p. 2). If true, it may become increasingly hard to infer what qualitative states exist in various synthetic entities; from AI-enhanced humans to laboratory-grown brain organoids, or advanced AI entities. Bostrom & Shulman further suggest these future advanced AI entities, which they term “digital minds”, will have justifiable moral claims, in virtue of their ability to have conscious experience: “Society in general and AI creators... have a moral obligation to consider the welfare of the AIs they create, if those AIs meet thresholds for moral status” (p. 3). The authors even consider the possibility that an advanced AI system could ‘create’ other, multiple digital minds, and abuse them, committing so-called “mind crimes” (p. 3). Whilst such ideas remain highly speculative, the idea of one AI system causing another AI system to ‘suffer’, so committing a mind crime, seems to take us very far beyond the scope and reach of the unitarians’ call for ‘equal consideration of equal interests’. It is not clear, as Singer’s theory demands, how one can start to establish the ‘equal interests’ of such an advanced AI entity that may or may not exhibit clear evidence of sentience or consciousness.

So, I have presented three problems for Singer’s unitarian theory of moral status, arising from whether equal interests can in fact exist and whether we can reliably know what these interests are; to what extent consciousness matters; and how well-equipped the theory is to ascribe moral status to future, synthetic entities. I will argue that Kagan’s hierarchical theory may be better suited to deal with these problems, but first I need to say more about how the hierarchical theory works.

4. The hierarchical theory

I have described how Kagan’s hierarchical approach considers the differential capacities of creatures for sentience, agency, psychological capacity (the capacity to enjoy a rich, complex mental life), and possibly consciousness. It also considers potential and modal conditions. Let me say a little more about each of these. To best describe the theory, I will jump forward to what Kagan describes as “the view that emerges” after he has proceeded through his

⁶ AI training computation is predominantly measured using floating-point operations per second, or ‘FLOPs’.

argumentation (Kagan, 2019, p. 292). This will explain the target theory for which I am arguing. I will then trace back my steps to consider questions and problems for the theory.

Kagan’s claim is that differing amounts of certain capacities, or ‘markers’ of moral status, permit us to reasonably ascribe differing levels of moral status to different entities. As previously noted, entities that are not and will never be sentient are not candidates for moral status. We could imagine these occupy ground zero, or ‘Level 0’ of moral status – entities at this level lack any moral status. Note that future intelligent machines do not necessarily fall into this category, as they might exhibit some form of sentience in the future. Any creature for which there is evidence of sentience, however minimal, will be deserving of some minimal level of moral status; let us call this Level 1. Above these might be creatures that have some psychological capacity and intentional agency, which might include a concept of the past, and wishes and desires for the future. This could include fish, birds, simple mammals, and a wide range of ocean-faring creatures, such as crustaceans. These might occupy Level 2. Above these might be all non-human animals that display complex psychological capacities, including pigs, dogs, octopuses, dolphins, and primates. Let them occupy Level 3. Sitting at the ‘top’ of this hierarchical tree would be normal humans, with their self-ascribed ‘full’ moral status, at Level 4. The intention of this sketched hierarchical example is not to claim any accuracy in ascribing a given creature to any given level, nor to make any claim as to the magnitude or multiple of the value of one level’s moral status over another, but rather to draw a plausible sketch of how moral status might be divided into banded hierarchies, based on sentience, agency and psychological capacities. For a summary of this hypothetical hierarchy, see **Figure 1**.

Level	Sentience, Intentional agency	Psychological capacities	Examples	Moral status
0	None	None	Rocks, stones	No moral status
1	Minimal	None/minimal	Sponge, oyster	Hierarchical levels of moral status
2	Moderate	Some	Birds, fish	
3	Complex	High	Dogs, dolphins	
4	Highly complex	Maximal	Normal humans	Full moral status

Figure 1: A hypothetical hierarchy of moral status

Here I will make a brief clarification on terminology. Some of the relevant literature refers to moral standing, as opposed to moral status, and sometimes interchangeably, e.g. Warren (1997). However, Buchanan (2009) claims we should distinguish between moral standing and moral status, claiming the former as a binary concept, whereas the latter term admits of degrees. To resolve this, I will consider *moral standing* to be that possessed by anything that has non-zero moral status. Moral status may admit degrees. In this way, I think the confusion falls away. I refer only to *moral status* in the remainder of this paper, for clarity.

The idea then, is that for any given creature, their membership of a group possessing ‘normal’ capacities for that creature determine that creature’s moral status. So, by way of an

example, all dogs should be regarded alike, in terms of their moral status: there is no differentiation across the entire range of creatures that are ‘dogs’. Stray dogs do not warrant some lesser moral status than, say, police rescue dogs. We might choose to treat police rescue dogs relatively better⁷, but morally, a stray dog matters ‘for its own sake’ as much as any other dog; it would be wrong to differentially harm a stray dog in comparison with any other dog. However, compared to an animal with markedly lower capacities, preferential treatment of a dog would be permissible. Kagan’s capacities approach permits the interests of a dog to be put ahead of, say, a mouse, or an insect. **Figure 2a** shows conceptually how this might look. A creature with minimal, but non-zero capacities (C_M), or ‘markers’ of moral status, might occupy Level 1. Dogs might exist at some higher level (e.g. Level 3). Humans are assigned full moral status (‘FMS’). **Figure 2b**, by contrast, shows the unitarian view, “where all levels of relevant capacities [above some minimal capacity C_M] result in the very same status” (Kagan, 2019, p. 283).

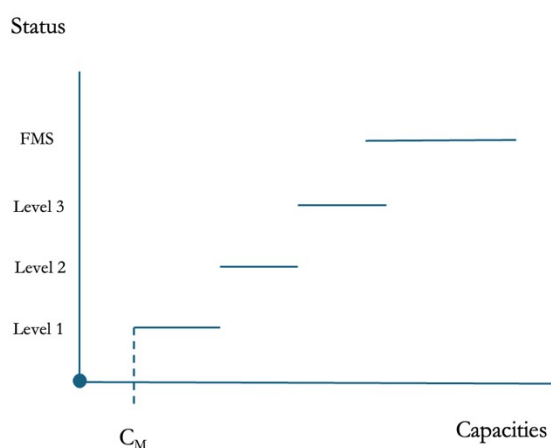


Figure 2a – the hierarchical view

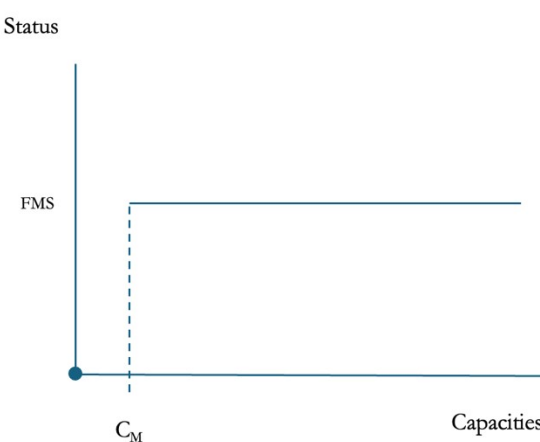


Figure 2b – the unitarian view

4.1 The importance of capacities

Earlier I described how the unitarian theory permits differential considerations where the ‘interests’ of creatures differ. So, you might ask, how does the hierarchical approach vary from or improve on the unitarian theory? The important feature is that Kagan’s theory targets *capacities*, which is a wider and richer concept than the ‘interests’ of Singer’s theory. Consider the following: it is possible that a mouse and a dog have an equal ‘interest’ in avoiding an equal pain: but this does not accommodate the significantly greater psychological capacities of the dog to interpret, remember, and fear future pain. The dog’s ‘psychological context’ is so much broader and deeper that its potential for suffering is significantly greater, such that it is deserving of a higher moral status than, for example, a mouse. So how plausible is it that a wide range of capacities are central to fixing differential levels of moral status? **Kagan clearly thinks so, as he lists the following extensive list of**

⁷ Presumably for utilitarian reasons: that we think police rescue dogs are ‘worth’ more, or ‘deserving’ of more, due to their contribution to society.

capacities: abstract and complex thought, creativity and imagination, conceiving the past and the future, making long-term, complex plans, self-awareness and self-consciousness, normative reflection and motivation, self-governance and autonomy, and individual choice.

(2019, pp. 126–127). It is variations in all these capacities, in level or sophistication, that according to Kagan lead to humans warranting greater moral status than animals, and some animals having greater moral status than others. It is notable though, that Kagan does not argue for a continuous, linear scale of capacities, mapping to a continuous scale of moral status. He does not suggest that some humans, purely in virtue of having certain greater capacities than others (such as better abstract reasoning, or physical strength), would be deserving of greater moral status – that would be problematic, just in the way that racism has been, and is still problematic. Nor would some ‘clever’ dogs warrant greater status than less clever dogs. Moral status seems to gather justifiably, according to Kagan, around a relatively small number of levels (2019, p. 293).

4.2 Potential and modality

A further feature of Kagan’s hierarchical capacities theory is that it incorporates the concepts of potential and modality in determining moral status. Consider a newborn baby. Its capabilities are limited; it can breathe, feed, cry and sleep, but not much else. In terms of pure capability at that time, it is wholly dependent on others and far less capable than an adult dog, chimpanzee or dolphin. However, it has the capacity to become a fully functioning adult human, and in almost all cases will do so. It has *potential*. It seems deserving of full moral status on account of its potential, indeed its likelihood, to be a fully functioning adult human. This could be considered *the argument for full moral status based on potential*. Consider now a severely cognitively impaired human, who for unfortunate reasons, maybe an accident in childbirth, will never be or become a normally functioning person. Many argue they should have the same moral status as a ‘normal’ human; it would be wrong to harm them in preference to a ‘normal’ human, they should not be treated unfavourably when compared to ‘normal’ humans. Crucially, we are saying that if that entity had turned out *how a typical example of that entity would have been*, given the *a priori* circumstances (in this case, two human, non-genetically compromised parents, a ‘normal’ conception, etc.), this individual could have, indeed would have, been a normal person. This could be termed *the argument for full moral status based on modality*. Incorporating potential and modality seems appealing, as it seems to line up a theory of moral status with our intuitions. Despite the capacities of a newborn baby being extremely limited, in human societies they are not treated as having inferior moral claims to a non-human animal of greater capability at that moment in time. It is the baby’s potential to become a full member of human society in the near future that grounds their moral status.

However, *the argument for full moral status based on modality* is sometimes used by unitarians to claim that the hierarchical view must be rejected. Their argument goes something like this: 1. We should not treat impaired humans worse than ordinary humans. 2. Some impaired humans have the same level of capacities as some animals. 3. So, those animals should be treated the same as ordinary humans. 4. Under the hierarchical account,

we do not treat animals the same as ordinary humans, so either the hierarchical account must fail, or premise 1 or 2 must be false. Since premise 2 is evidently true, we must either reject premise 1 and treat impaired humans worse; or the hierarchical account must fail (Kagan, 2019, p. 157). However, Kagan rejects this argument, and I think he is right to do so. Kagan argues that the impaired human, if certain conditions about the world had been different, could have been a normal human. It is this possibility, in a counterfactual world, that grants the impaired human the possibility of full moral status. Importantly, those ‘differences about the world’ are taken to be those that would have *normally* and *usually* occurred. He is arguing that *if the statistically normal set of outcomes* for that class of relevant entity had occurred, the individual could have, would have been a normal human. By contrast, an animal, even one functionally equivalent to an impaired human, could *never* have been a normal human – and so does not qualify for the same moral status as normal (or impaired) humans.

This interpretation may, however, leave a residual problem for Kagan to address. How it can be that, due to modality, a cognitively impaired human is awarded full moral status, but in certain cases of comparing the *value* of the life of a severely cognitively impaired human with a normal human, for example, to allocate scarce life-saving treatment, or to choose who to save in a conjoined-twins problem, it might be permissible to differentiate? This is a difficult problem to resolve, as in matters of practical ethics, we are sometimes forced to make that choice, and rule in favour of what might be termed the ‘normal’ human, or at least the human with the greatest life-chances ahead of them. Kagan accepts this as true, but makes the following important observation: just because in some limited scenarios, we accept that a cognitively impaired human has a lower moral status than a normal human, this *does not entail* that some cognitively impaired humans might have a status equal to or lower than that of any animal (2019, p. 161). It is consistent with a hierarchical account that, whilst some cognitively impaired humans might have a lower status than normal humans, *all* cognitively impaired humans can claim a status above the highest non-human animals, in virtue of their modal status. To recap, as there might be a hint of a contradiction here. The hierarchical account, or “the view that emerges”, as described by Kagan (2019, p. 292), admits different levels of moral status between groups with meaningfully different capacities, as shown in **Figure 2a**. The modality of this account permits, for some moral purposes, impaired members of a group, members that *could have reasonably been fully functioning members* of that group, to be considered as moral equals. An example would be, in minimising trivial harms to either, no preference should be given. However, the account also permits *differentiation within group members* who have *meaningfully different capacities*, where those morally relevant features are important to necessary ethical judgements. This might include life or death choices, where a choice must be made. This account still permits cognitively impaired humans to *always* have a moral status higher than that of any non-human animal, on account of modality.

4.3 Two worries for Kagan’s modality

I'd now like to discuss two worries for Kagan's theory that arise from its use of modality in determining moral status. The first arises from, under Kagan's modality, how group membership is determined; and the second comes from the possibility of unintuitive outcomes. For a creature such as a cognitively impaired human to claim full moral status because they 'could have been' a normal human, presupposes that some group membership exists. But where does this claim to group membership come from? If the answer is species membership, then this theory of moral status is open to the charge of "speciesism" – that it merely favours members of its own species unreasonably, in the same way that sexism or racism favours members of the same gender or race for no good reason. According to unitarians such as Singer, this would violate a fundamental moral principle, that of "equal consideration of interests", as already discussed (2023, p. 5). Kagan, however, rejects that the hierarchical theory is guilty of speciesism, arguing that it might be rather a case of "modal personism" (2015, p. 17). Kagan refers to "someone who is not a person but who *could* have been a person...as a '*modal person*'" (Kagan, 2019, p. 138). I have already explained Kagan's concept of modality. The concept of 'person' represents those creatures that can become, or are, 'normal' humans. The idea here is that a biological human, whose genetic makeup is such that their 'genetic peers' would normally go on to be persons, qualify for full moral status. What Kagan proposes, is that, whilst many – including Kagan – agree with Singer it is probably wrong to be "speciesist", it is not necessarily wrong to be a *modal personist*, in virtue of the merits of the hierarchical account of moral status.

But what if the group membership, to which Kagan pins modality, turns out to be identical to the group that is defined by species? After all, a species is generally defined as those which can and do reproduce together, creating genetically similar copies of themselves. It sounds rather like the group of 'genetic peers' I described above. Kagan goes to some length (2019, pp. 13–20) to establish what defines a relevant group, and how that is not necessarily equivalent to species – although I am not sure that he fully succeeds here. But maybe a better defence can be found elsewhere. Simply put, Kagan argues that it is the *capacities* (or "relevant features") of individuals that make up a group count, not *membership* of the group itself (2019, p. 118). So, according to this account, if we encountered a Martian dog, that had all the psychological and physical capacities as an Earth dog, we would probably afford the Martian dog the same moral status as an Earth dog – even if it turned out to have a completely different genetic makeup to Earth dogs, and in no way could qualify as a member of the Earth-bound species group *Canis lupus familiaris*. So presumably, a synthetic, silicon entity that exhibited *identical capacities* to a normal human would count for full moral status, even if it were clearly nothing like *Homo sapiens*. I will not further consider this speculative possibility here, but I'll just note that what grounds group membership seems to be an important question for the success of the hierarchical view.

A second worry is that of the possibility of unintuitive outcomes. Consider this example posed by Kagan: the case of *infant anencephaly*, a condition where the upper part of the brain is missing (Kagan, 2015, p. 16). Imagine a case of two infants born alive, but with anencephaly. Case 1 is caused by a genetic defect: there is *no way* that that child could have developed normally, given their genetic makeup. Case 2 is caused by, say, environmental poisoning *in utero*; they 'could have' developed differently, if their environment had been

different. According to Kagan’s modal theory, Case 2 would be a modal person, but Case 1 is not – yet this doesn’t seem intuitively right, that they should be treated differently, from a moral point of view. If we are to consider Case 1 to be a ‘modal person’, we seem to need additional flexibility to extend the concept of modality. As Kagan acknowledges, “a great deal turns on controversial issues in the metaphysics of modality” (2015, p. 16).

So I have highlighted two worries for Kagan’s idea of modal personhood that underlies the hierarchical theory of moral status. Unitarians might insist these are significant worries. However, I think that a strength for Kagan is that his theory lines up rather well with our intuitions; that actual, potential or modal capacities all seem to track what we commonly take to be the markers of moral status in considering who is deserving of moral regard, and at what level.

5. Reasons to prefer Kagan’s theory

I have criticised the unitarian account for its reliance on a fact of the matter that equal interests exist; on our (in)ability to *know* the ‘interests’ of other creatures, given the difficulties in detecting sentience and of *knowing* about the nature of other minds. But in what regard, if at all, does the hierarchical account outperform the unitarian here? The hierarchical theory may be equally bound by the limits of neuroscience, and by our conceptual tools in recognising the interests of, or sentience in others. Kagan acknowledges the necessary imprecision of this task, and so proposes an approach to assigning moral status under the hierarchical theory, which he refers to as “realistic rule consequentialism” (2019, p. 285). Rule consequentialism is the ethical approach whereby the greatest good is sought by following a set of rules that, if generally followed, will lead to the best outcomes, or consequences. It can be contrasted with ‘act consequentialism’, which requires choosing the action, however complex or counterintuitive, that maximises the desired consequences in that situation⁸. One advantage of rule consequentialism is that it accommodates epistemic uncertainty; that is, it does not require perfect knowledge of which particular act maximises the good in any particular situation. In the context of assigning moral status, realistic rule consequentialism accepts that we cannot know for certain the capacities and sense experience of other creatures. Rather, we can take a *realistic*, as opposed to *ideal*, approach to applying rules that are likely to match the reasonable assignment of moral status to capacities. An ‘ideal’ approach, by contrast, would require “all moral agents to conform to the relevant rules, perfectly and without mistake” (2019, p. 285). As this can be seen as overdemanding, realistic rule consequentialism is preferred. For Kagan, the hierarchical view to be applied is a ‘realistic’ approach to evaluating moral rules, as it accepts our epistemic limitations. Kagan terms this approach “practical realism” (2019, p. 292).

I have also criticised unitarians for putting the difficult problem of consciousness ‘at the front of the reasoning chain’ in establishing necessary criteria for moral status. How

⁸ These explanations are more commonly made with respect to *utilitarianism* (as opposed to consequentialism), but as we are interested in well-being here, they are similar enough not to spend time discussing differences that are of no matter.

might the hierarchical approach avoid this criticism? Kagan’s approach looks first for signs of sentience, but also at the capacities, including psychological capacities, and potential of a creature or entity. It considers any ‘modal’ claim to moral status of an entity. It can take a precautionary approach – as recommended by Birch. In this way, Kagan’s hierarchical, capacities approach is more likely to award moral status to a creature not deserving of it, than to deny it to a creature that should have it. But this seems the right way to fall, given what is at stake. In doing so, it pushes the ‘hard question’ of consciousness to the ‘end’ of the reasoning chain, as it were, by which time it has already carried out the difficult, important moral philosophical work of determining moral status.

In **Section 3.3** I criticised Singer’s unitarian view for being somewhat poorly equipped to deal with novel situations that we may encounter soon, such as AI-enhanced humans or fully synthetic, AI-enabled agents. How might Kagan’s hierarchical theory better cope with these? Earlier I covered the possibility of synthetic entities being created with similar, or superior capacities to humans. Without speculating on the likelihood of this occurring, it is worth spending a moment considering the implications. As mentioned, Bostrom & Shulman consider the prospect of future advanced AI entities having superhuman moral claims, either through having stronger morally relevant interests, termed “super-beneficiaries”, or through having greater moral status, presumably from their greater psychological capacities, termed “super-patients” (2023, p. 3). The super-beneficiary is a plausible extension of a thought experiment originally presented as a problem for utilitarianism by Nozick (1974, p. 41), where he proposed the “utility monster”. Nozick’s monster is a creature that gains so much utility from a given resource, that their interests are prioritised above all others. If such superhuman moral claims were to come about, what is to stop these claims from being at or possibly beyond that of full moral status? This would conceivably put humans in the position that intelligent non-human animals are in today; animals are granted some rights, but due to their inferior technological and intellectual power, they are beholden to how humans choose to treat them. One response to this slightly uncomfortable prospect is as follows. Some have suggested that ‘full’ moral status is in fact the maximal moral status that can exist (Kagan, 2019, p. 150). In this case, even future super-capacity beings would have a moral obligation to treat humans equally. Could such a limit exist? Whilst there might be a finite limit on concepts such as memory or computational power, it is less clear that there should be limits on concepts such as “normative reflection... or the ability to grasp abstract truth” (Kagan, 2019, p. 151). So, there may be no good reason to assume that such a boundary just happens to coincide with where human moral status is today. A further possibility is that creatures superior to us would be superior in their moral outlook, and so might choose, both rationally and morally, to act with benevolence toward us. Furthermore, the extent to which humans are involved in their design and programming might increase this possibility. But it is not at all clear what will transpire. Where we are left, then, is as follows: such advanced creatures might exist in the future; Kagan’s hierarchical account of moral status could readily accommodate such novel creatures, based on their relevant capacities; and we probably have no good reason to deny that they might have higher moral claims than us. What seems to count for the hierarchical theory is that its architecture can start to address, in an incremental fashion and by investigating

capacities, the question of the possible moral status of machines, without a commitment to either consciousness, or the limits to moral status, on which we can only speculate.

6. Concluding remarks

I have argued that Singer’s popular, unitarian account of moral status, whilst lining up with our simple intuitions about pain and suffering, faces three problems. These problems centre around the existence and knowledge of equal interests in unequal creatures; its overreliance on determining consciousness, and its applicability for future synthetic entities. In all these regards, I think Kagan’s hierarchical theory offers a superior account, given its ability to incorporate a wider range and concept of capacities, especially psychological capacities, that entities might possess. Kagan’s incorporation of potential and modality is also a valuable feature, but as I have shown, this can run into problems of defining relevant groups, and it may produce unintuitive outcomes. I have considered some difficulties in evaluating the importance of consciousness to moral status, and favour identifying markers of moral status ahead of determining consciousness in an entity. It seems pragmatic to establish moral status through Kagan’s capacities approach ahead of tackling the significantly harder questions of consciousness. Turning to the question of machines and moral status, Kagan’s theory offers an appealing intellectual infrastructure. The capacities approach can be applied quite readily to the functional features of a machine. A final thought, however, is that in applying the theory’s modality, this might prompt a bifurcation in how we come to view moral status: between, on the one hand, biological creatures who originated on Earth; and on the other, our future silicon cousins, who could conceivably be functionally equivalent or superior to us; but in the sense of relevant group membership, *never* like us. It might be that we endorse and retain a hierarchical capacities approach to the moral status of biological creatures, but alongside this emerges a separate, parallel hierarchy, equally considerate of the interests, capacities or ‘inner lives’ of synthetic entities, but somehow not commensurate with it. It is too early to say how we will make comparisons or choices across such hierarchies, but more work in this direction may be of interest. One approach to this problem of commensurability might be to consider the social relations that emerge between biological and synthetic entities, and to give these greater weight in our moral assessments. But that is a further project. As things stand, I believe I have shown that Kagan’s hierarchical, capacities approach has much to be said for it, and further work in this direction should be pursued, as we head towards a more technologically enabled future.

Bibliography

Bentham, J. (1789). *An introduction to the principles of morals and legislation*. T. Payne and Son.

- Birch, J. (2024). A walk along the edge. In *The edge of sentience: risk and precaution in humans, other animals, and AI* (Oxford, 2024; online edn, Oxford Academic, 19 July 2024) <https://doi.org/10.1093/9780191966729.003.0002>
- Bostrom, N., & Shulman, C. (2023). *Propositions concerning digital minds and society* (Version 1.21) [Manuscript in preparation]. Forthcoming, 2025, in *Cambridge Journal of Law, Politics, and Art*. <https://nickbostrom.com/propositions.pdf>
- Bryson, J. J. (2010). *Robots should be slaves*. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues* (pp. 63–74). John Benjamins Publishing Company. <https://doi.org/10.1075/nlp.8.11bry>
- Buchanan, A. (2009). Moral status and human enhancement. *Philosophy & Public Affairs*, 37(4), 346–381. <https://doi.org/10.1111/j.1088-4963.2009.01164.x>
- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
- Crisp, R. (2025). Well-being. *The Stanford Encyclopedia of Philosophy* (Fall 2025 Edition), Edward N. Zalta & Uri Nodelman (eds.), forthcoming. <https://plato.stanford.edu/archives/fall2025/entries/well-being/>
- DeGrazia, D. (2021). An interest-based model of moral status. In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking moral status* (pp. 39–55). Oxford University Press. <https://doi.org/10.1093/oso/9780198821892.003.0002>
- Epoch AI. (2025). *Parameter, compute and data trends in machine learning* [Dataset]. Retrieved April 24, 2025, from <https://epoch.ai/data/notable-ai-models>
- Financial Times. (2025, July 30). Inside the relentless race for AI capacity. *Financial Times*. <https://ig.ft.com/ai-data-centres/>
- Garland, A. (Director). (2014). *Ex machina*. [Film]. A24
- Harman, E. (2003). The potentiality problem. *Philosophical Studies*, 114(3), 173–198. <https://doi.org/10.1023/A:1025381802116>
- Jaworska, A., & Tannenbaum, J. (2023). The grounds of moral status. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 edition). <https://plato.stanford.edu/archives/spr2023/entries/grounds-moral-status/>
- Kagan, S. (2015). What’s wrong with speciesism. *Journal of Applied Philosophy*, 33(1), 1–21.
- Kagan, S. (2019). *How to count animals, more or less*. Oxford University Press.
- McMahan, J. (2021). Suffering and moral status. In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking moral status* (pp. 22–38). Oxford University Press. <https://doi.org/10.1093/oso/9780198821892.003.0002>

- Metzinger, T. (2009). *The ego tunnel: The science of the mind and the myth of the self*. Basic Books.
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.
- Nozick, R. (1974). *Anarchy, state, and utopia*. Basic Books.
- Papineau, D. (2025, forthcoming). Consciousness is not the key to moral standing. In G. Lee & A. Pautz (Eds.), *The importance of being conscious*. Oxford University Press.
- Parfit, D. (1984). *Reasons and persons*. Oxford University Press.
- Rawls, J. (1999). *A theory of justice* (Rev. ed.). Belknap Press of Harvard University Press.
- Schukraft, J. (2020, May 18). *Comparisons of capacity for welfare and moral status across species*. Effective Altruism Forum.
<https://forum.effectivealtruism.org/posts/4Xg3dC6rrW4WFnSne/comparisons-of-capacity-for-welfare-and-moral-status-across>
- Singer, P. (2023). *Animal liberation now*. The Bodley Head.
- Singer, P. (2025, June 12). *The ethical crisis lurking in the background of the AI boom* [Interview]. Big Think.
<https://bigthink.com/series/the-big-think-interview/sentient-ai-peter-singer>
- Szücs, A., Varona, P., Volkovskii, A. R., Abarbanel, H. D. I., Rabinovich, M. I., & Selverston, A. I. (2000). Interacting biological and electronic neurons generate realistic oscillatory rhythms. *NeuroReport*, 11(3), 563–569. <https://doi.org/10.1097/00001756-200002280-00027>
- Warren, M. A. (1997). *Moral status: obligations to persons and other living things*. Oxford University Press.