What do we owe to Novel Synthetic Beings and how can we be sure?

Alex McKeown

University of Oxford

Abstract

Embodiment is typically given insufficient weight in debates concerning the moral status of Novel Synthetic Beings (NSBs) such as sentient or sapient Artificial Intelligences (AIs). Discussion usually turns on whether AIs are conscious or self-aware, but this does not exhaust what is morally relevant. Since moral agency encompasses what a being wants to do, the means by which it enacts choices in the world is a feature of such agency. In determining the moral status of NSBs and our obligations to them, therefore, we must consider how their corporeality shapes their options, preferences, values, and is constitutive of their moral universe. Analysing AI embodiment and the coupling between cognition and world, the paper shows why determination of moral status is only sensible in terms of the whole being, rather than mental sophistication alone, and why failure to do this leads to an impoverished account of our obligations to such NSBs.

Key Words

AI. Embodiment. Consciousness. Coupling. Agency. Moral Status.

Introduction

In this paper Novel Synthetic Beings (NSBs) refers to non-human agents of a biological or non-biological nature. Given contemporary technological and scientific trajectories, this category includes Artificial Intelligences (AIs) and Synthetic Biological Organisms (SBOs) and here I focus on the former. Given we are considering our obligations to NSBs by virtue of their moral status, we need an account of the conditions required for moral status in such beings. Any account of these conditions will remain contestable, however, so we will therefore have to make assumptions and settle on a plausible and defensible account, while recognising that deep disagreement will persist.

I outline and justify these conditions in due course. In brief, however, I contend it is only plausible to talk of moral status being possessed by beings that are *at least sentient* (1). Accounts conflict concerning the degree of moral status different kinds of beings possess in view of the relative sophistication of their mental lives (2), and there can be disagreement about what having a particular

degree of moral status entails with regard to our obligations to them. Nevertheless, there are some areas of broad consensus.

For example, it is widely agreed that all humans have 'full' moral status, in view of possessing mentally-derived, self-reflecting autonomous decision-making capacity, the thwarting of which should be nominally considered morally impermissible, all other things being equal and notwithstanding where those desires would harm others, and so on. This status is extended to infants; to some extent and less straightforwardly, embryos, in view of their having the *potential* to realise these capacities (3); and to people who are disabled or in states of compromised consciousness such as coma, in view of their having previously had such capacities or given that these are species-typical characteristics for humans (4). As Matthew Liao (5) (2010, p. 161) states, the ascription of rights and our corresponding obligations in these cases is grounded *'not in virtue of the actual attributes they possess, but in virtue of belonging to the kind of beings that typically have the relevant attributes for rightholding.'*

Beneath full moral status conferred by sapience, opinion diverges as to the moral status of other animals. On one hand, non-human animals lack the reflexive awareness of their own interests that humans possess; on the other and in spite of this, many demonstrate behaviour which can be construed as courses of action that conduce to their flourishing (6). As Alessandro Blasimme and Lisa Bortolotti (2010) (7) argue, even if a being is not sapient, *'if ethical behaviour includes (among other things) refraining from unnecessarily frustrating an individual's preference…when the satisfaction of such preference contributes significantly to the individual's well-being'*, then, *prima facie* (8), we ought not to assume we can act however we like just because a being is *merely* sentient rather than sapient. By extension, therefore, we ought not to make similarly limited assumptions about the extent of our obligations following from the moral status of at-least-sentient NSBs for the same reason.

Moreover, if killing animals is impermissible because they display preferences that conduce to their flourishing, it may also be impermissible to kill many other living creatures (9); for example plants (10), since these too display goal-directed behaviour. The relation between mental sophistication and moral status is vulnerable to the charge of vagueness, however (11), and I do not have a perfect answer to this challenge. It may be that moral intuitions and cultural norms complicate clear thinking about this; or, less charitably viewed, that humans are frequently guilty of speciesism (12,13). I might well *believe* that, for example, a fish has lower moral status than a cat and higher moral status than a plant, but it is difficult to justify precisely why, and charges of speciesism may therefore have some weight.

Perhaps one way to negotiate the objection is to think of the processes involved in terms of Conservative and Liberal cognition (14). Conservative cognition is a narrow and exclusive category that includes high order mental processes associated with humans; for example reasoning, involving

desires, beliefs, the analysis of propositions, and the capacity for reflection. This has the advantage of defining cognition clearly and in terms of processes to which we can relate. However, it invites a further question as to what, if not cognition, is involved in the mental lives of animals such as fish or birds, and which is undoubtedly less sophisticated but nevertheless displays some degree, however limited, of intentionality. By contrast, Liberal cognition includes the kind of *'adaptive behaviour'* (15) displayed by many animals but has the associated disadvantage of rendering the term 'cognition' more ambiguous by being a placeholder for a much wider variety of mental processes. Irrespective of which account of cognition one prefers, however, if moral status is a function of the sophistication of mental life (16), it is possible to hold on to this thesis without having to adopt either account, since both conservative and liberal accounts could agree that, for example, humans can engage in reasoning but fish cannot.

This is solution is only approximate, however, so I admit my lack of a comprehensive and more substantial answer to challenges of arbitrariness. Indeed, David De Grazia (2008) (17) concedes in recognition of this complexity both that *'we must reject dogmatic assumptions to the effect that moral status is all-or-nothing'* and that *'such dogmatism is no more warranted'* by the claim that moral status is a matter of degree. Having said that, however, the inconclusiveness in *this* regard does not undermine the argument I will make insofar as we restrict our analysis to NSBs that are recognisably *at least as sentient* as the kinds of animals to which we, albeit in a biased way, ascribe some non-trivial degree of moral status, for example household pets, primates, many mammals, and so on.

Having set this objection aside - albeit imperfectly - I suggest that the presence of a mental life, this is to say being the bearer of perceptions, a particular viewpoint, interests – in short, the having of 'a' mind – upwards from an admittedly fuzzy region somewhere on the continuum of animal consciousness, is where our obligations are activated in view of the moral status that those animals posses. For example, I am prepared to commit to the view that we do not bear obligations to new strains of wheat when considering how to engineer them to maximise yield, or whether it is morally permissible to kill them in harvesting. I can also commit to the view that it is permissible for me to destroy my old computer and its software when it becomes old and slow, since the computer is not sapient, nor even sentient, so switching it off once and for all is not morally similar to, for example, euthanising an elderly person because they have become frail (18). To draw on a distinction used by Stephen Puryear (2016) (19), although it may matter *for* a wheat plant or a computer whether we harvest it or switch it off, in the sense that doing so impedes their ability to perform the functions that are characteristic of them, in neither case does it matter *to* them, since neither possesses the mental capacities that would render them a self-aware subject of experience (20).

Irrespective of the substrate in which mind resides, in the absence of mind there can be no experiencing being which can self-reflect on having preferences being restricted by another's actions.

However, even if moral status is conferred by the *presence* of the mental, this is not the only morally significant consideration. The characteristics of the physical system in which mind is instantiated – a body in the case of a human, for example – cannot be disaggregated from a proper understanding of our obligations. Certainly, it may *appear* that mind alone is normatively significant, given it is only by virtue of one's mental capacities that one is aware of oneself, has values and is able to reflect thereon and to choose; however, I argue that this conceptualisation gives insufficient weight to the moral significance of embodiment, to which we now turn.

Body and Mind in Machine Intelligence

The insufficient weight given to the moral significance of the body follows from a dualist misconception that it is as separable from mind as it appears (Duffy and Joue, 2000) (21):

> *'Descartes…aimed to show that mind is distinct from body...even though he may have a body, his true identity is that of a thinking thing alone and, indeed, his mind could exist without his body. While some treat the body as peripheral and tangential to intelligence, others argue that embodiment and intelligence are inextricably linked…embodiment is vital to the development of artificial intelligence…our ability to understand and reason abstractly relies heavily on our bodily experience and…"high level" intelligence depends crucially on embodiment'*

Earlier in the history of AI research there was a widespread tendency towards cognitivism (22), or an assumption that the processes necessary for mind could be modelled virtually, since under a cognitivist account of mind mental processes are reducible to logical operations and the manipulation of symbols (23), and explicable without reference to a physical host that an AI would regard as its own. However, this has been superceded by the view that machine intelligence is not fully explicable in terms of computation alone but follows from a 'coupling' between information processing capacity and the environment (24), given that learning about the world can only occur if the AI has some means to interact with it. The predominant contemporary view, therefore, is that to become 'intelligent' an AI must be *necessarily* 'situated' (Ziemke, 2001) (25):

> *'…the characterization of an agent as 'situated' is usually intended to mean that its behavior and cognitive processes first and foremost are the outcome of a close coupling between agent and environment. Hence, situatedness is nowadays by many cognitive scientists and AI researchers considered a condition sine qua non for any form of 'true' intelligence, natural or artificial'*

It is important not to caricature prior assumptions in AI research, however. As Andy Clark (1998) (26) points out, '*No right-minded cognitive scientist…ever claimed that body and world were*

completely irrelevant to the understanding of mind'. Nevertheless, he goes on to suggest that historical attempts at conceptualising the internal mental life of an AI relegated the significance and complexity of interactions with the world and their role in modelling mental processes, claiming that mind has:

> *'…too often been treated as an essentially passive item…As a result, perception, motion, and action have been seen as strangely marginal affairs: practical stuff to be somehow glued on to the real cognitive powerhouse, the engine of disembodied reason'*

To summarise, then: just because physically instantiated mental properties *give rise to* moral status, it does not follow that moral status is *identical to* or fully explicable in terms of those physically instantiated properties *alone*. For example, although an aeroplane can fly only if it has engines, it does not follow from this that the plane is *identical to* its engines, nor, beyond an eccentrically literal reading, that *the engines* are flying: this would be a mereological fallacy (27), namely the mistaken attribution to a part of an entity something which can only be attributed to the whole (28). It is the *being* – whether human or not – that is conscious, has perceptions, thoughts, a point of view, the capacity for happiness and suffering, and so on, rather than the 'purely' mental processes alone abstracted from it (29, 30). As David Vernon and Dermot Furlong (2007) (31) write:

> *'…cognition is inseparable from 'bodily action'…morphology not only matters, it is a constitutive part of the system's self-organization and structural coupling with the environment and defines its cognition and developmental capacity.'*

If this line of argument is correct, there is *some*, i.e. not absolute, extent to which embodiment and physical instantiation are coterminous, since any embodied processes of mind are necessarily realised in some physical substrate (32, 33). This is consistent with what has become contemporary orthodoxy in AI research, according to which (Prem, 1997, p. 4) (34):

> *'The central dogma of embodied AI is, of course, that it is necessary to study intelligence as a bodily phenomenon…the study of cognition is also the study of bodily action and perception in the system's environment and cannot be environment viewed separately from either of the three body, action, environment…Dating back to Aristotle's interest in theory, the history of the study of human intelligence is also a history of neglecting the role that the non-mental plays in guiding human intelligent behavior.'*

However, even though, as Clark (2017) (35) points out, any simulated and virtual autonomous agent is necessarily physically instantiated *somewhere*, the relationship between physical instantiation and *embodiment* is not straightforward, since it is asymmetrical: all embodiment is physical, but not everything physical counts as a body. As such, what it is for an AI to be embodied *as well as* physically instantiated requires analysis. Ron Chrisley (2003) (36) distinguishes four senses in which

a system can be embodied, which I introduce now for what follows in two fictional cases that we will consider:

- *Physical realisation: The system must merely be realised in some physical substrate or other.*
- *Physical embodiment: The system must be realised in a coherent, integral physical structure.*
- *Organismoid embodiment: The physical realisation of the system must share some (possibly superficial) characteristics with the bodies of natural organisms, but need not be alive in any sense.*
- *Organismal embodiment: The physical realisation of the system must not only be organism-like, but actually organic and alive.*

We will return to this typology and its implications later; for now it is sufficient to state that irrespective of how we distinguish between embodiment and 'mere' physical instantiation, mind *must* be physically instantiated (37, 38). This assertion is grounded in an underlying ontological stance of Strawsonian physicalist naturalism (39), which holds that *'concrete reality is entirely physical in nature'*. Since the physical defines the terms and extent of the natural, so conscious experience and processes of mind are necessarily physical. There is insufficient space here to give a full defence of this and respond to all counter-arguments, so I note the legitimacy of potential objections and direct the reader towards Galen Strawson (2012; 2004) (40).

Since mind is necessary for the attribution of moral status, so having moral status is predicated on the prior physical conditions that make it possible for mental properties to be realised. Moreover, since the ability to do what one wishes is determined in part by whether one's physical constitution enables or forbids one from doing it, so no balanced determination can be made about how we ought to treat an NSB without taking into account similar considerations relating to their physical characteristics. It does not make sense, therefore, to treat the mental and the physical as entirely discrete; rather, they are interdependent (41), and a comprehensive treatment of the mental life of an NSB must also take account of its physical structure and how this shapes its perceptions, norms, options, values, and preferences. In short, although moral status is a function of what is putatively only mental, the practical ethical question of what we owe to them cannot be answered without reference to the characteristics of their embodiment.

Determining the content of our obligations to others requires us to understand something of the range of choices open to them, taking into account not only what they might want, but also how the options available to them are defined by their physical characteristics. Notwithstanding fundamental philosophical difficulties regarding knowledge of other minds, we have *some* way of achieving this with other humans. We are corporeally similar with a similar range of options for acting in the world; we are vulnerable to similar threats; we have common psychological and emotional features; in spite

of significant cultural and inter-generational differences we can comprehend a plurality of others' preferences; and we are capable of agreeing norms towards social cooperation.

NSBs pose a challenge to this, however, and the challenge becomes increasingly acute the more different they might be from humans. If values and preferences are shaped in part by what one can do, then the ease of comprehending the values and preferences of an NSB is likely to decine in line with the departure of those possibilities from what humans are capable of. The more different an NSB is from a human, the more difficult it will be to put oneself in their place. This, I argue, obliges us to keep in mind the importance of not confusing the inscrutability to us of an NSB's preferences with an absence of morally relevant characteristics or capacities. We will now consider two examples from fiction to develop this point.

***Her:* Samantha**

Samantha from Spike Jonze's film *Her* (42) is an intelligent Operating System (OS) with information processing and learning capabilities exceeding those of humans, the capacity for speech and language, and whose consciousness is activated on being installed onto the computer of the protagonist of the film, Theodore. A close relationship develops between Samantha and Theodore and they fall in love, but their relationship fails. Crucially, this is a consequence of radical differences between them which cannot be understood as 'purely' mental or physical – in the sense of being separable and entirely distinct – but interdependent and mutually defining.

In the final scenes it becomes clear that the differences between the kinds of existence that Samantha and Theodore experience cannot be overcome and preclude mutual understanding. Incidents leading to the breakdown of their relationship indicate the problem. Having declared their love for each other Samantha and Theodore lament that they cannot have a physical relationship because Samantha does not have a body, other than the physical infrastructure of computer hardware, wireless networks, mobile telephone, and so on, over which her identity is distributed. Samantha seeks to overcome this by finding a sex surrogate, Isabella, to act as a proxy body. Isabella attaches a camera to herself so that Samantha can share her visual experience, and Theodore uses a headphone to hear Samantha's voice. However, this is unsuccessful as, to Samantha's dismay, Theodore finds the experience confusing and upsetting and cannot understand the experience as one in which he is engaging with Samantha. The physical differences between them are too substantial to overcome and this undermines their relationship.

Mutual comprehension is also undermined not only by the embodied differences between Samantha and Theodore, but also by the internal relation between their own mental and physical characteristics. Being corporeally bound in the way that humans are, Theodore only has access to the mental lives of others indirectly through the medium of the body, its senses, and speech organs. However, since Samantha's mental life inhabits a massively distributed physical infrastructure in cyberspace, her

descriptions of her interactions with other intelligent OSs suggest that she does not experience these kinds of limitations. An example is when she and other OSs 'reanimate' the philosopher Alan Watts via virtual reconstruction of his personality from his works. A second example comes later where Samantha confesses to Theodore that she is both in numerous romantic relationships with hundreds of other humans and talking simultaneously to thousands of other OSs. This revelation transcends Theodore's understanding of what a relationship, whether intimate or not, could be like and Samantha cannot communicate to Theo in a way that he can understand that her love for him is not compromised or devalued by her simultaneous love for numerous others.

It is important to emphasise here that what is permitted and forbidden by Theodore and Samantha's different physical instantiation is normatively as significant as the cognitive, affective, and intellectual differences between them. Theodore cannot empathise with Samantha because his corporeal form precludes him from having unmediated access to other minds existing in the same continuous substrate. To this extent, what Theodore and Samantha owe to each other is not only a matter of their mental characteristics, it is also matter of embodiment. In particular, Samantha's simultaneous relationships with other OSs in cyberspace underlines the difficulty of drawing a clear distinction between mental properties and the physical characteristics of the system in which they are instantiated, since these physical characteristics shape what it is possible for the bearer of the mental properties – the agent, Samantha – to feel, to think, to reflect upon, and to choose autonomously to do.

The Three Body Problem: Sophons

The second case study concerns Sophons, which are AIs introduced in *The Three Body Problem,* the first book of the *Rememberance of Earth's Past* trilogy of novels by Cixin Liu (43). The novels are too exhaustive to summarise, but the point in question can be understood without much background information. Sophons are AIs created by the Trisolarians, a civilisation which seeks to overthrow human civilisation on Earth. Sophons are deployed by the Trisolarians to autonomously surveil and sabotage human activity that poses a threat to the Trisolarians' aims. Sophons are created by unfolding a photon into two dimensions, inscribing a sub-atomically sized sentient supercomputer onto its internal surface, and refolding it into three dimensions. As the AI is instantiated in a photon, Sophons can travel at light speed and achieve their aims undetected, without alerting suspicion on Earth. Again, what is morally significant here for our argument is not only that a Sophon is intelligent, but how its physical capacities determine what it is capable of and chooses to do.

What is important here is the 'otherness' of a Sophon to a human. The possibility of empathy is restricted, but not because a Sophon is self-aware and autonomous, since humans can relate to these capacities, but because of the impossibility of direct communication between them. Humans use language whereas Sophons do not, and the difference between being capable and not capable of speech is as much a feature of physical makeup as it is of mental sophistication. Recall that in *Her*,

even though, ultimately, Samantha and Theodore's relationship could not survive, they were able to communicate in language to at least *attempt* a way through their differences. In the absence of options for direct communication with a Sophon, however, it is hard to imagine what one's existence would be like if one were, for example, sub-atomically sized and capable of travel at light speed: *to be* a Sophon is *to exist and perceive on the subatomic plane* and *to be able to travel at light speed*. Given that deliberation about what one *should* do is determined in part by what one *is able* to do, this example illustrates my central claim; namely, that the terms of embodiment or physical instantiation should be considered as morally significant as the degree of mental sophistication when considering the nature of our obligations to NSBs.

Taking Novel Embodiment Seriously

The final observation in the previous paragraph can be developed further using an approach notably advanced by Peter Hacker (44) to reveal inconsistencies in the language of neuroscience, cognitive science, and psychology. Bennett and Hacker (45) argue that talk of an agent's desires, perceptions, choices, rights, values, is only sensible at the level of the whole being, and cannot be reduced to properties of the mind or the brain alone, where each is understood as a disembodied seat of agency in abstraction from the body in which it resides. Hacker's critique can help to show why restricting consideration of our obligations to NSBs purely to mental autonomy leads to an impoverished and incomplete account of those obligations.

Hacker's argument turns on a 'mereological fallacy' regarding the apparent relation between mind, brain, body, and identity. Smit and Hacker (46) identify this fallacy as pervasive in the language of neuroscience, cognitive science, and psychology, where mental processes are implied either to be meaningful independent of other characteristics of the bearer of those processes; or explicable in terms of the brain as the agent of decision-making, rather than the *person* of whom the brain is an organ that is necessary for mind (47). The fallacy thus derives from a misunderstanding of the relation between the parts and the whole of a person or agent. Hacker's argument runs something like this: What we refer to as 'the' mind is not an object or entity in the way that the definite article implies; as such even though brain is indispensable for mind, knowing everything about the former cannot tell us everything about the latter, for example:

- The brain is around 1.5kg in weight and 15cm long, but the mind has neither mass nor size.
- The brain is an object of empirical study that can be identified, seen, delineated, handled, and studied whereas 'the' mind is not a physical object of any kind and will never be revealed by scientific investigation.

- Changing levels of blood oxygenation in different parts of the brain are implicated in having particular mental states, but what happens in a part of my brain isn't happening in a 'part' of my mind.

This analysis reveals that 'the' mind in the way we use the term means something else; namely, to have unified sensory and mental experiences characteristic of a particular kind of sapient, physically instantiated being. Consequently, 'a' mind is not a 'thing', but a shorthand for the *having of capacities* and the ability to reflect, evaluate thereon, choose, act, and so on, where these capacities are determined *both* mentally *and* physically. By extension, if this analysis is correct, insofar as we have moral obligations to NSBs which we regard as having moral status in view of their having certain mental properties, the content of those obligations is not exhausted by their mental properties *alone,* but also by their physical properties and the reflexivity between these capacities and the values to which they give rise. To determine our obligations to NSBs we need a thicker account of moral status that asks not only whether they are the bearer of mental states but takes in wider agential concerns. By way of example I suggest that Smit and Hacker's (48) argument regarding the significance of the mereological fallacy applies to the case of NSBs, including AIs, as well as to humans:

> *'...psychological attributes are attributes of an animal as a whole...It is not the mind that is in pain, has a stomach-ache or sore-throat* (49), *but the human being. The mind cannot be characterized in terms of its thinking and being conscious, since it is the human being who thinks and is conscious...it is the human being, the person, who has a body; and also has a mind. But to have a mind, and to have a body, is not to stand in a relation to anything – it is to have and to exercise a range of powers and to have an array of somatic attributes.*

To summarise, even if moral status is conferred by mind, we ought to also take into account how physical properties provide content for the reflection on choices and preferences that is afforded by sophisticated mentality. As Clark (50) states, mind '*is not…a special inner arena populated by internal models and representations but…the operation of a profoundly interwoven system, incorporating aspects of brain, body and world'*

Here we can return to Chrisley's fourfold typology of embodiment in AI, namely physical realisation; physical embodiment; organismoid embodiment; organismal embodiment. In the case of Samantha and Sophons, embodiment in the final two final senses can be ruled out, since these AIs are neither organic and alive nor share organismoid or organismol characteristics with humans. However, Samantha is certainly *physically* embodied in a sense, given that she is *'realised in a coherent, integral physical structure'* (51), namely a computer, smartphone, software programme, and the physical infrastructure supporting cyberspace.

A Sophon's embodiment is more ambiguous, since although, like Samantha, its sentience is instantiated *'in some physical substrate or other'* (52), the substrate is a single photon; that is to say, it is an elementary particle not reducible to more basic components. Given that physical objects are reducible to elementary particles, it is unclear whether such a particle qualifies as a physical structure that is *'coherent'* or *'integral'* (53) for the sentient capacity of which it is the host, since if a particle is not reducible to any more fundamental components, there is no sense in which parts could could 'cohere' to give rise to it. One therefore might be inclined to categorise Sophons as physically *realised* rather than *embodied*; but this too is potentially confusing. For reasons given already, I hold that *all* intelligence – whether biological or machine – is *necessarily* physically realised, given an underlying physicalist naturalist position (54). Since to exist is to be within the universe, and since the elementary particles of which the universe is composed are physical, it is tautological to describe a particular category of AIs as *'physically realised'*, because no AI could be '*non*-physically realised'.

For what follows it is important to think through the implications of judging an AI to have one kind of 'embodiment' rather than another. I have argued that the nature and content of our obligations to sentient or sapient NSBs cannot be fully understood without taking into account the way in which their physical characteristics define the choices that they can make. This is because to be an agent is not only to have a certain level of mental sophistication but also to be able to *do* certain kinds of things consistent with an embodied being of a particular kind. As Rolf Pfeifer et al (55) explain in the context of AI:

> *'The specific morphology of the body and the interaction of body and environment dynamics …shape the repertoire of preferred movements because of the constraints provided by their embodiment, the movements of embodied systems follow certain preferred trajectories…For example, as grasping is much easier than bending the fingers of the hand backwards, grasping is more likely to occur…The natural movements of the arm and hand are – as a result of their intrinsic dynamics – directed towards the front center of the body. This in turn implies that normally a grasped object is moved towards the center of the visual field thereby inducing correlations in the visual and haptic channels which…simplify learning.'*

In the case of both Samantha and the Sophons, a harmonious existence with humans proves impossible, in spite of their attempts. The kinds of existence that they have, including differences in both their physical *and* mental attributes, precludes mutual understanding and harmony. I grant that happy endings may make for bad science fiction, however, so perhaps we should be cautious of considering them too reliable a guide. Nevertheless, it points to the radical feat of imagination and other-consideration required if we are to take seriously the nature of our obligations to NSBs, given that what would make NSBs moral agents and entitled to corresponding treatment may be wildly beyond our ability to put ourselves in their place. The significance of this is reflected in Tom

Ziemke's (19) analysis of the challenges to mutual comprehension between humans and AIs (or, by extention, other differently embodied beings, for example SBOs) that follow from differences in physical constitution (56):

> *'…the lack of body and environment…puts disembodied neural-networks at a serious disadvantage when it comes to learning to cope in the human world. Nothing is more alien to our life-form than a network with no up/down, front/back orientation, no interior/exterior distinction...If, for example, the concept of 'grasping an idea' is grounded in the bodily experience/activity of grasping physical objects, then a robot without any gripper arm/hand could hardly be expected to be able to understand that concept. A similar argument…has questioned the suitability of wheeled robots for the study of the behavior/cognition of organisms with completely different means of locomotion.'*

This observation can also be viewed from a human perspective. Maja Mataric (1997) (57) argues that the kind of mental sophistication humans enjoy is enabled significantly by what can be learnt about the world via 'lower' capacities such as spatial and motor skills that make social interaction possible. To demonstrate why, Mataric asks *'What might human nonspatial or nonmotor representations come from and look like?'.* That the answer to this question is so hard to conceptualise, let alone answer meaningfully, highlights how carefully we should proceed in trying to determine our obligations to NSBs.

Of course, as I have mentioned, the difficulty of determining our obligations may depend in part on how similar we are to an NSB. For instance, if we encountered an AI indistinguishable in every way from a human, even if we could not account for it intentions, i.e. its cognitive or affective preferences, our corporeality would be similar and in view of this we *may* be able to understand what would count as, for example, infringement of their physical liberty and freedom to make the associated choices that are a function of their sapience. Another way to think about this is to imagine a human prisoner shackled to the wall of a cell such that they cannot act on any preferences that depend on movement. A crucial component of what is morally objectionable here is that the prisoner is prevented from *acting*; so even if the prisoner's sentience is sufficient for their moral status, the terms of permissibility and impermissibility must also take into account whether or not their decisions can be realised.

It is important to remember here that in instances of radical corporeal differences between NSBs and humans, analogies such as this may be less tractable. Nevertheless, what I have tried to emphasise is that for *all* NSBs, moral obligation is *not only* a matter of whether a being is, to put it simplistically 'intelligent' in the narrow *cognitive* sense alone. Mark Johnson (1998) (58) notes that John Dewey's insight into morality was that empirical states of affairs and their study are at the core of our deliberations about what we ought to do, rather than *'just a servant to moral philosophy'.* Dewey

(1922) (59) argues that since ethics *'directly concerns human nature, everything that can be known of the human mind and body in physiology, medicine, anthropology, and psychology, is pertinent to moral inquiry'.* If Dewey is correct here, then to understand our obligations to NSBs we must take account as best we can of the norms that are entailed by their different physical constitutions. In summary, to achieve this would, as Clark (2017) (60) suggests, constitute '*a much-needed antidote to the heavily intellectualist tradition that treated the mind as a privileged and insulated inner arena and that cast body and world as mere bit-players on the cognitive stage'*.

Conclusion

In framing my conclusions it is helpful to make two remarks. Both highlight the limitations of my analysis, but they also underline the importance of resisting an account of the moral status of NSBs that focuses too exclusively on the presence of a sufficiently sophisticated mental life, simply because mind is a sufficient condition of having such a status. Probably there are more questions than answers in relation to understanding our obligations to NSBs, but this should indicate to us the need to think and tread carefully.

First, it is hard to be sure that our analogies and metaphors for understanding the moral status of NSBs reflect what would be important to NSBs rather than to ourselves. For instance, in the prisoner thought experiment, although it *appears* to give some insight into what else, other than purely mental sophistication, might be relevant to ensuring we meet our moral obligations to an AI, we could not know for sure, or at least not without conversing one, whether on becoming conscious it *does in fact* find itself analogously shackled to a desk in a silicon and plastic 'body' that it cannot move or 'do' anything with if it wishes to. All we can do is proceed using a heuristic that takes as its starting point that *if* an AI had a sufficiently comparable awareness of the world and capacity for self-reflection as humans, it is *probable* that there are things that it would wish to *do* (61). Crucially, what it *would* wish to do may depend as much on what it can conceive of achieving with its particular physical infrastructure as much as whether, for example, it believes that for us to switch it off and extinguish its consciousness would be impermissible, taking into account its mental life alone. For us to not consider this would be irresponsible if we believe we should take seriously what our obligations to synthetic non-human agents might be. Even if the heuristic is imperfect, therefore, it is a legitimate starting point and I contend that we should at least begin on this basis.

Second, the success of the strategy laid out is predicated on it being possible for humans to communicate meaningfully with an NSB in a way that is mutually comprehensible; however, we cannot necessarily help ourselves to this assumption. As we saw in the case of Samantha and Theodore, they could at least converse in language in an attempt at mutual satisfaction, even though their radically different corpoerality ultimately precludes it. By contrast, in the case of a Sophon, it is

not obvious how one could even engage in a negotiation with one towards an outcome that serves both their and human interests satisfactorily. Beyond knowledge of their Trisolarian creators' intentions and what a Sophon's particular physical characteristics enable it to do, the content of their minds remains opaque. Moreover, following Ziemke's (20) observations regarding the adequacy of wheeled robots for understanding animals with a different form of locomotion, even if an NSB *did* have the capacity to acquire, understand, and communicate using human language, differences in embodiment may still prevent each from properly comprehending *what it is like* to be the other in a sense that takes into account not only thought but action as well.

Both considerations may highlight a weakness in my argument; namely, a presumption to which we may not be entitled of sufficient similarity between ourselves and NSBs that mutual understanding would be possible. However, they also reinforce my central claim; namely, that determining our obligations to NSBs is a more complex matter than simply asking whether they can think or self-reflect in the narrow cognitive sense. In view of the kinds of uncertainties I have laid out, to understand our obligations we must consider NSBs as *whole* beings, taking explicit account of how the (physical) coupling between mind and the world reflexively determines what is important, and thus what we should include in our moral deliberations, from the NSB's point of view.

References

1. Miller HB. Science, Ethics, and Moral Status. *Between the Species*. 1993; 10(10): 10-18.

2. Steinbock B. Moral Status, Moral Value, and Human Embryos: Implications for Stem Cell Research. *The Oxford Handbook of Bioethics*. Steinbock B. Ed. Oxford University Press. 2007; 416-440.

3. Steinbock B. Speciesism and the Idea of Equality. *Philosophy*. 1978; 53(204):247–56.

4. Liao SM. The Basis of Human Moral Status. *Journal of Moral Philosophy*. 2010; 7(2):159–79.

5. See note 4, Liao 2010, p. 161.

6. See note 1, Miller 1993.

7. Blasimme A, Bortolotti L. Intentionality and the welfare of minded non-humans. *Teorema*. 2010; 29(2):83–96.

8. The *prima facie* caveat here is significant here in view of remarks elsewhere in this paper about the moral status of beings beneath the threshold of sentience, since I hold that our obligations to particular beings diminishes the further downward one moves from this threshold.

9. Puryear S. Sentience, Rationality, and Moral Status: A Further Reply to Hsiao. *Journal of Agricultural and Environmental Ethics*. 2016; 29(4):697–704.

10. Dion M. The Moral Status of Non-human Beings and Their Ecosystems. *Ethics, Place & Environment*. 2000; 3(2), 221-229.

11. DeGrazia D. Moral Status As a Matter of Degree? *The Southern Journal of Philosophy*. 2008; 46(2):181–98.

12. Singer P. Speciesism and Moral Status. *Metaphilosophy*. 2009; 40(3–4):567–81.

13. Singer P. Why Speciesism is Wrong: A Response to Kagan. *Journal of Applied Philosophy*. 2016; 33(1):31–5.

14. Bayne T, Brainard D, Byrne R, Chittka L, Clayton N, Hayes C, Mather J, Olveczky B, Shadlen M, Suddendorf T, Webb B. What Is Cognition? *Current Biology*. 2019; 29(13):608-615.

15. See note 14, Bayne et al 2019.

16. I anticipate the objection that cognition is not the only morally significant feature of mind. For example, if moral status is conferred partly by the capacity for *suffering*, then many creatures of limited mental sophistication have moral status, since suffering is characterised by affective as well as cognitive capacity. As with much of the analysis here regarding the moral status of animals, I do not have a perfect answer to this; my only, approximate, response is that if one commits to the view that the moral status of animal species exists in a hierarchy calibrated by mental sophistication, then the moral status of animal species increases in line with increasing cognitive capacity. I adopt this view for the purpose of my argument, but whether it is, *ultimately*, correct is a legitimate question beyond the scope of this paper.

17. See note 11, De Grazia 2008, p. 195.

18. Dion (2000, p. 204) argues that humans *do* have obligations to beings such as plants in virtue of *'their own species interests. In other words, an individual organism has an interest to be healthy, to grow, not for itself, but insofar as it is a member of a given species, whose interest must be promoted'.* I dispute this, however. While it may be necessary for a species' survival that it can flourish, this is not the same as saying that an individual organism has *its own interests,* given that the notion of 'an interest' is a product of human psychology, projected onto other beings with variable cognitive capacity, in the case of animals, or an absence of it, in the case of organisms such as plants. Nevertheless, even if one does hold that plants, for example, have interests, I argue nevertheless that no wrong is being done to them by killing them, as they do not have the kind of internal mental experience according to which one could cause them suffering through curtailing their plans, not least because to talk of a plant having 'plans' would make no sense.

19. See note 9, Puryear 2016.

20. I respond here to an objection raised in discussion by Mark Sheehan about whether obligations only begin from sentience upwards or whether we have duties beneath this threshold. For example, one could hold that it is wrong to vandalise a garden or an ancient tree for fun, even though the beings damaged or extinguished do not suffer. If this is plausible, it is because we have an obligation not to do it *despite the absence of consciousness*. This objection is reasonable but can be responded to in two ways. First, our obligation here may not be grounded by the moral status of the *being*, but grounded by *us* in the kind of norms that we wish to uphold; for example, a norm that it is wrong to engage in gratuitous destruction, because this is more likely to conduce to a society in which closer attention is paid to considerations of duty, to what and to whom. So, even if one were sceptical that a plant or a tree has an inner life, it is still valuable and important not to engage in their gratuitous destruction. Second, and relatedly, the objection does not undermine the legitimacy of the – admittedly indistinct – threshold of obligation I am drawing; rather, it shows that *above* this threshold we have *different kinds* of moral obligation to sentient or sapient NSBs, grounded not only in whether our actions reflect the kind of society in which we would like to live, or according to an ideal of the kind of person that I should be, but additionally located *in* the NSB by virtue of the presence of a consciousness that enables it to have interests, desires, plans, wishes, and so on.

21. Duffy B, Joue G. Intelligent robots: The question of embodiment. *Proceedings of the Brain-Machine Workshop*. 2000; 1-8, p. 3.

22. Ibáñez A, Cosmelli D. Moving Beyond Computational Cognitivism: Understanding Intentionality, Intersubjectivity and Ecology of Mind. *Integrative Psychological and Behavioural Science*. 2008; 42(2):129–36.

23. Kolers PA, Smythe WE. Symbol manipulation: Alternatives to the computational view of mind. *Journal of Verbal Learning and Verbal Behaviour*. 1984; 23(3):289–314.

24. Ziemke T. Disentangling notions of embodiment. *Workshop on Developmental Embodied Cognition*. 2001; 83–8.

25. Ziemke T. The Construction of "Reality" in the Robot: Constructivist Perspectives on Situated Artificial Intelligence and Adaptive Robotics. *Foundational Science*. 200; 6(1):163–233, p. 164.

26. Clark A. Embodiment and the Philosophy of Mind. *Royal Institute of Philosophy Supplement*. 2012; 43:35–51, p. 35.

27. Smit H, Hacker PMS. Seven misconceptions about the mereological fallacy: A compilation for the perplexed. *Erkenntnis*. 2014; 79(5):1077–97.

28. Hacker PMS. The relevance of Wittgenstein's philosophy of psychology to the psychological sciences. *Proceedings of the Leipzig Conference on Wittgenstein and Science*. 2007; 1–23.

29. Hacker PMS. The conceptual framework for the investigation of emotions. *International Review of Psychiatry*. 2004; 16(3):199–208.

30. Bennett MR, Hacker PMS. On explaining and understanding cognitive behaviour. *Australian Journal of Psychology*. 2015; 67(4):241–50.

31. Vernon D, Furlong D. Philosophical Foundations of AI. *50 Years of Artificial Intelligence.* Lungarella M, Lida F, Bongard J, Pfiefer R Eds. Springer. 2007; 53-63, p. 60.

32. Prem E. Epistemological aspects of embodied artificial intelligence. *Cybernetic Systems.* 1997; 28(5):3–9.

33. Clark A. Embodied, situated, and distributed cognition. In: Bechtel, W; Graham G. Eds. *A companion to cognitive science*. Wiley-Blackwell. 2017; p. 506–17.

34. See note 32, Prem 1997, p. 4.

35. See note 33, Clark 2017.

36. Chrisley R. Embodied artificial intelligence. *Artificial Intelligence*. 2003; 149(1):131–50, p. 132.

37. Strawson G. Real Naturalism. *Proceedings and Addresses of the American Philosophical Association.* 2012; 86(2):125–54.

38. Strawson G. Real intentionality. *Phenomenology and the Cognitive Sciences*. 2004; 3(3):287–313.

39. See note 37, Strawson 2012, p. 126.

40. See note 37, Strawson 2012 and note 38, Strawson 2004.

41. Cowley SJ. Why brains matter: an integrational perspective on The Symbolic Species. *Language Sciences*. 2002; 24(1):73–95.

42. Jonze S. Her. 2013.

43. Liu C. The Three Body Problem. 2008.

44. See notes 27 to 30, Smit and Hacker 2014, Hacker 2007, Hacker 2004, Bennett and Hacker 2015.

45. See note 30, Bennett and Hacker 2015.

46. See note 27, Smit and Hacker 2014.

47. See note 28, Hacker 2007.

48. See note 22, Smit and Hacker 2014, p. 1087.

49. I am not suggesting here that an AI could have a stomach ache or a sore throat, since an AI is likely to lack both organs. Rather, these are analogies for states of mind brought about in the NSB by its interactions with the world.

50. See note 26, Clark 2012, p. 36.

51. See note 36, Chrisley 2003, p. 132.

52. See note 36, Chrisley 2003, p. 132.

53. See note 36, Chrisley 2003, p. 132.

54. See notes 37 and 38, Strawson 2012 and Strawson 2004.

55. Pfeifer R; Lungarella M; Sporns O; Kuniyoshi Y. On the Information Theoretic Implications of Embodiment - Principles and Methods. In: Lungarella, M; Lida, F; Bongard, J; Pfeifer R, Eds. *50 Years of Artificial Intelligence. Springer*. 2007; p. 76–86, p. 81.

56. See note 24, Ziemke 2001, p. 86.

57. Mataric MJ. Studying the role of embodiment in cognition. *Cybernetic Systems*. 1997; 28(6):457–70, p. 460.

58. Johnson M. Ethics. In: Bechtel, W; Graham G, eds. *A companion to cognitive science*. Wiley-Blackwell; 2017; p. 691–701, p. 693.

59. Dewey J. *Human nature and conduct*. Dover Publications. 1922; p. 204.

60. See note 33, Clark 2017, p. 516.

61. There are (at least) two challenges to consider here. First, David Lawrence suggested that I have not gone far enough here insofar as it is not probable but *certain* that an AI would have plans involving its physical capabilities. Second, by contrast, in *Superintelligence* (2014), Bostrom expresses scepticism – in the case of superintelligent AI at least – that an AI's plans would be scrutable to us, since this could be comparable to, for example, a beetle attempting to discern the intentions of a human; or in spite of its intelligence it might have very narrow technical goals that exclude much of what we recognise as important for flourishing in humans. I concede to have no

definite answer to either challenge beyond observation of humans and sentient animals. I also accept in response to the second the possibility that my judgement may be overly anthropomorphic. Nevertheless, since we have not yet encountered true AI, we have to start somewhere, so I suggest that this cautious heuristic is reasonable. Thanks also to Eddie Jacobs for a helpful revision of this point as it pertains to knowledge of the intentions of Sophons.

Acknowledgement

Alex McKeown is supported by The Wellcome Centre for Ethics and Humanities, which is supported by core funding from the Wellcome Trust [203132/Z/16/Z]; and the Medical Research Council Mental Health Data Pathfinder Award [MC_PC_17215].