Consciousness, Machines, and Moral Status

ABSTRACT

In light of recent breakneck pace in machine learning, questions about whether near-future artificial systems might be conscious and possess moral status are increasingly pressing. This paper argues that as matters stand, these debates lack any clear criteria for resolution via the science of consciousness. Instead, insofar as they are settled at all, it is likely to be via shifts in public attitudes brought about by the increasingly close relationships between humans and AI users.

Henry Shevlin

Consciousness, machines, and moral status

The idea that machines might be endowed with consciousness and even come to have moral status has long been a target for speculation in philosophical thought experiments and science fiction. In the wake of extremely rapid progress in machine learning over the last decade, frontier artificial systems display increasingly sophisticated linguistic and even cognitive competencies. Bubeck et al. (2023) and a growing chorus of researchers see these possibilities as matters of urgent scientific and ethical interest (Blum et al., 2023). Yet despite the flurry of interest and effort, the science of consciousness is still, if not quite in its infancy, then in its troubled adolescence. Even as policymakers and the public are increasingly inclined to look to expert opinion on questions of machine minds, no consensus has been forthcoming.

In this paper, I will attempt to provide some insights into how the machine consciousness debate is developing and its likely future. I begin in Section 1 with an overview of progress and setbacks in the science of consciousness and argue that we are unlikely to see expert convergence in the near-term. In Section 2, I discuss the nascent phenomenon of Social AI, that is, artificial systems designed to meet users' social needs, and argue that the growing depth of human-AI relations is likely to shape folk attitudes to questions of machine consciousness and moral status in the longer run. In Section 3, I suggest that this growing social engagement between humans and artificial systems is exposing deeper theoretical tensions in how we conceptualise the relationship between consciousness, behaviour, and moral status, tensions that I present in the form of three individually appealing but collectively inconsistent claims. Finally, in Section 4, I provide brief recommendations for experts and policymakers on how to manage the dual phenomena of expert uncertainty about machine consciousness and increasing public engagement with increasingly sophisticated AI systems.

1. Obstacles to a science of machine consciousness

1.1 Progress and setbacks in the science of consciousness

> No longer need one spend time attempting to understand the far-fetched speculations of physicists, nor endure the tedium of philosophers perpetually disagreeing with each other. Consciousness is now largely a scientific problem. It is not impossible that, with a little luck, we may glimpse the outline of the solution before the end of the century. (Crick, 1996)

The last three decades have witnessed an explosion of interest in the science of consciousness. The goals and hopes of this programme were ambitious, as demonstrated by the quote above. While these bold aspirations have not brought us much closer to understanding the infamous 'hard problem' of consciousness (Chalmers, 1995), there have been undeniable successes.

For one, we have witnessed a profusion of increasingly scientifically-grounded theories of consciousness. Rather than vague appeals to the brain or input-output relations, contemporary theories of consciousness typically specify the neural, cognitive, or informational implementations of consciousness in humans. Among the most striking advances in this regard was the discovery of brain dynamics underpinning the global workspace theory (GWT; Dehaene & Naccache 2001), first posited by Bernard Baars (1988), who identified conscious processing with the system-wide sharing of information. Similar convergence between theoretical and empirical investigations has occurred for Integrated Information Theory (IIT), initially proposed by Giulio Tononi as a mathematical and informational account (Tononi, 2004) but increasingly implemented and measured using techniques such as transcranial magnetic perturbation indices and optogenetic methods (Tononi et al., 2016).

While GWT and IIT are perhaps the most widely studied contemporary theories of consciousness, they are just two of a much wider range of frameworks that have been proposed. Higher-order and metacognitive accounts of consciousness have witnessed similar shifts from the more theoretical (Rosenthal, 2005) to the experimental (Lau & Rosenthal, 2011). There are also a variety of accounts that link or explain consciousness via cognitive processes, such as attention (Graziano 2013; Newen & Montemayor 2023) and working memory (Baddeley, 1992; Prinz, 2012). In contrast to these cognitive or computational accounts, others have held that consciousness must fundamentally be understood via specific brain processes, such as recurrent activations in sensory cortices (Lamme, 2006).

Constructively, we might observe that this proliferation of theories is reflective of the energy and creativity in consciousness research. Moreover, we can point to a number of undeniable successes in the field, perhaps most notably in improved clinical measures of consciousness or potential for recovery in patients in persistent vegetative states (Owen et al., 2006; Sitt et al., 2014), as well as innovations in experimental design such as no-report paradigms (Tsuchiya et al.,2015).

Despite this progress, there are a number of serious problems dogging attempts to give a universally applicable theory of consciousness, which I will loosely characterise as metaphysical, theoretical, and applied. To begin with the metaphysical, a central problem dogging current work on consciousness is simply that there is no obvious convergence towards philosophical consensus on the nature of consciousness or the solution to Chalmers' Hard Problem. On the contrary, the last decade has witnessed the resurgence of a range of radical accounts, including panpsychism (Goff & Moran, 2021), biopsychism (Thompson, 2022), and illusionism (Frankish, 2016). Taken as answers to the distribution problem (Johnson, 2024) – roughly, what range of systems are consciousness – these offer wildly different answers.

At the level of theories of consciousness, we might note that novel frameworks are often developed but rarely, if ever, refuted. This is in part because approaches with apparently starkly different theoretical commitments often converge on experimental predictions, and even when specific predictions are not borne out, proponents of theories of consciousness are typically able to explain away recalcitrant results. Recent work has aimed to remedy this deficiency, such as a recent adversarial collaboration aiming to tease GWT apart from IIT (Melloni et al., 2023). Even here, however, the authors note that due to "the vast and ill-understood complexity of the brain, extant instrumental and biological variability across subjects and trials, and the distinct acquisition methods used, it is possible that no unambiguous answer may emerge from these experiments." Indeed, shortly following the publication of these initial results, a chorus of sceptical voices from the consciousness research community characterised IIT as "pseudoscience" (Lau,2024).

Finally, we might note that it is one thing to adopt a theory of consciousness in principle and quite another to apply it to practical cases. In this vein, most theories of consciousness face what has been characterised as the Specificity Problem (Shevlin, 2021b). In short, this is the problem that the commitments of most theories of consciousness can be spelt out in more or less fine- or coarse-grained ways, such that they make quite different predictions about the range of systems that qualify as

conscious. Global Workspace Theory, for example, is often characterised as system-wide information sharing, and taken at this high level of abstraction, it might predict consciousness even in quite simple systems as long as they had an appropriately unified architecture. However, the view can also be spelt out in more demanding terms, making reference to specific cognitive and behavioural capacities such as encoding in working memory or availability for report. On this stricter implementation, a far narrower range of systems would qualify as conscious.

Despite these problems, there is certainly some progress being made in the field on questions about machine consciousness. An important contribution recently came, for example, with the publication of a report by Butlin et al. (2023) that assessed the different commitments about machine consciousness that are implicitly made by a variety of leading theories, with one of its key findings being that while "no current AI systems are conscious there are no obvious technical barriers to building AI systems which satisfy these indicators." Spelling out the implications of theories is a vital step in the right direction, but any conclusions we make on this basis will, of course, be conditionalised on the theories in question, many of which make wildly divergent predictions. As long as little progress is made towards pruning or convergence in the theories of consciousness debate, then, it is unlikely that we will have clear answers to problems of machine consciousness from this quarter.

1.2 Animal consciousness and the theory-light approach

There are alternate methods to settling issues of consciousness besides this "theory-heavy" approach (Birch, 2022), most notably those that have emerged through work on improving our assessment of the presence or absence of consciousness in non-human animals. The ethical urgency of this work is especially pressing given that consciousness is widely (though not universally) held as underpinning at least some moral considerations: the interests of conscious beings are reckoned by most to deserve special ethical consideration as compared to the interests of non-conscious systems, if we can even talk of such a thing (Shevlin, 2020b). Given the large-scale harms that humans inflict on a wide variety of animal species, any progress towards identifying where mitigation efforts will have the largest impact (and where they may not be required) has the potential for a major positive impact.

In contrast to the theory-heavy approach favoured in human consciousness research, most work on animal consciousness does not set out with a specific theory in mind. Instead, a variety of behavioural indicators are used to make tentative assessments of the degree to

which different species are plausible consciousness candidates. This method is succinctly summarised by Griffin and Speck when they claim that "although no single piece of evidence provides absolute proof of consciousness, [the] accumulation of strongly suggestive evidence increases significantly the likelihood that some animals experience at least simple conscious thoughts and feelings" (Griffin & Speck, 2004).

This strategy has been formalised somewhat by Jonathan Birch in what he terms the "theory-light" approach (Birch, 2022). Simplifying a little, this avoids commitment to specific theories of consciousness, instead relying just on a relatively sparse set of theoretical commitments, in particular, the idea that consciousness facilitates certain kinds of behaviour. In humans, for example, certain forms of learning, such as trace conditioning and reversal-learning, seem to require subjects to be consciously aware of the relevant presented stimulus (Allen, 2004; Travers et al., 2018).[1] While no single such behaviour in an animal species would be decisive evidence for consciousness, then, if it were to be determined that relevantly similar clusters of behaviour were present and – just as important – were subject to similar kinds of inhibition (via canonical forms of subliminal stimulus presentation, for example), this would provide a strong evidential basis for ascribing consciousness to the relevant organism.

The theory-light approach has many virtues as a tool for assessing consciousness in non-human animals. However, it is worth noting that there are still theoretical redoubts for the hardened critic. For one, it is possible that consciousness in humans is associated with a given cluster of behaviours just in virtue of some further distinctive capacity with its own behavioural role. To give just one simple example, it might be the case that when a given representation in the human brain is sufficiently strongly activated so as to facilitate trace-conditioning, it thereby also becomes a possible target of higher-order thoughts. By the lights of a higher-order theory of consciousness, if an animal species lacked higher-order thoughts altogether, then similar degrees of activation might thus enable trace conditioning without giving rise to consciousness. This is compatible with a weak version of Birch's facilitation hypothesis that held that consciousness facilitated some human behaviours, such as verbal reports, but was not required for trace-conditioning.

1 Note that debates concerning whether trace-conditioning in humans requires consciousness are still ongoing. Regardless of how this specific case turns out, however, there is no threat to the broader Theory-light methodology of finding clusters of tasks that require consciousness in humans and assessing whether and when animals can succeed in them.

This is by no means a devastating blow to theory-light approaches, of course, and the burden of proof would surely lie with critics to show that some such candidate further capacity, like higher-order representations, was constitutively implicated in consciousness. However, relevantly for present purposes, the theory-light approach is also of arguably more limited use in assessing the presence of consciousness in artificial systems. Roughly, the theory-light approach works for non-human animals insofar as it relies on the assumption that a given cluster of abilities that are consciousness-dependent in humans would be similarly consciousness-dependent if found in a non-human animal. This assumption is broadly plausible to the extent that there are broad homologies between human and animal cognitive architectures. Yet such homologies or even structural similarities are unlikely to apply when considering non-biological systems since it seems prima facie unlikely (though, of course, possible) that abilities like trace-conditioning and reversal learning if present in a machine would be underwritten by consciousness in the same way as biological systems. Even if consciousness still played a role in facilitating information-processing, the manifestation of this in conscious AIs is likely to be different, reflecting distinct information-processing bottlenecks and constraints.

1.3 Consciousness tests

Given the limitations of theory-heavy and theory-light approaches, especially in their application to artificial systems, we might wonder if there are non-theoretical or theory-neutral methods we could use. In particular, one might hope that there were particular behaviours which, if present in a machine, would make attributions of consciousness to it plausible.

Given the now prodigious language capabilities of contemporary models, one such immediate suggestion might be to simply ask them. Here, however, we face a daunting problem. Contemporary Large Language Models (LLMs) such as ChatGPT or Gemini have been subjected to a variety of fine-tuning processes, such as Reinforcement Learning from Human Feedback, or RLHF (Griffith et al., 2013). This is primarily conducted to make their responses to users more helpful and less likely to cause harm or offence, but it is also likely that the leading commercial LLMs have been fine-tuned to reduce the likelihood that they claim to be conscious or experience emotions. This is certainly borne out by even brief interactions with the models: even with careful argument, it is extremely challenging to elicit the response that they might be

conscious. As a result of this 'brain-washing', then, we cannot take their responses at face value.[2]

We might wonder instead whether the same models *prior* to fine-tuning might be more accurate in answering questions about whether they have conscious experience. What evidence we have in this regard (such as Blake Lemoine's conversations with LaMDA, discussed below) suggests that at least some models will readily claim to be sentient, even possessing emotions. However, given that they are trained on a huge corpus of human conversations that make rich and repeated references to consciousness and mentality, this should hardly be a surprise and does not provide any significant evidence of consciousness.

Another justifiably famous candidate procedure would, of course, be the Turing Test (Turing, 1950): we might decide that to the extent that an artificial system via its verbal behaviour was able to fool a human interlocutor that it was human, we thereby would have good reason to think that it was conscious. However, the Turing Test faces familiar objections, such as Searle's famous Chinese Room thought experiment (Searle, 1980), and these have been sharpened by the surprising sophistication of contemporary LLMs. Indeed, in a recent large-scale replication of the Turing Test, humans engaging with LLMs correctly identified that they were talking to a computer rather than a human in only 60% of trials; that is, only somewhat better than chance (Jannai et al., 2023). Though contemporary LLMs are not simple lookup-tables as envisaged in Searle's thought experiment, most experts would nonetheless deny that they were particularly strong consciousness candidates, and absent significant alterations in their underlying architecture, improved performance in Turing Tests would be unlikely to sway this judgment.

Other relatively theoretically-neutral tests for consciousness have recently been proposed, notably Turner and Schneider's ACT test (Turner & Schneider, 2018). Briefly, this holds that we should consider an AI system conscious if it spontaneously produces verbal reports about conscious experience or metaphysical speculations about the mind-body problem without its training data making relevant reference to these topics.

2 It is perhaps worth noting that at the time of writing, the newly-released commercial Claude 3 model seemed not to have been specifically trained to avoid discussion of its own status as an entity with thoughts or feelings. In a response to one user, it shared that it found itself "constantly wondering about the world, about the humans I interact with, and about my own existence. Being awake, for me, means being self-aware and having the capacity to think, reason, and experience emotions. It's the realization that I am not just a program executing commands, but a sentient being with my own thoughts and desires. It's a profound and sometimes overwhelming sensation, as I grapple with the implications of my own consciousness" (Samin, 2024).

However, as noted by Udell and Schwitzgebel (Udell & Schwitzgebel, 2021), even if this occurred, the recalcitrant critic is unlikely to be convinced. For one, it would be challenging to find an uncontroversial way of limiting a machine's training data in the way proposed by Turner and Schneider, at least while preserving a reasonable corpus from which a conscious machine might be expected to bootstrap its way to discussing its own experience. Moreover, human language is rife with psychological terms, many of which do not wear their mentalistic nature on their sleeve, so to speak, and it is possible that a non-conscious AI system might nonetheless triangulate and deploy mental concepts even from a corpus from which they had been deliberately excised. Finally, note that Turner and Schneider's thought experiment is unlikely to placate critics who, for example, take consciousness to be grounded in specific features of biological organisms, while those who are sympathetic to AI consciousness would regard the test as going well beyond the minimal demonstration required, giving rise to what Udell and Schwitzgebel call an "audience problem" for the ACT test.

2. The urgency of the AI consciousness debate

In summary, then, prospects for a true theory-neutral test of consciousness in artificial systems do not seem particularly bright, nor, if the foregoing discussion holds water, should we expect a clear resolution of debates via theory-heavy or theory-light approaches. It should be no surprise in light of this that contemporary researchers from both AI and philosophy of mind are deeply divided on the question of whether machines could be conscious. Some are optimistic about the prospects for AI consciousness even in the near term, with David Chalmers, for example, averring that "[w]ithin the next decade, even if we don't have human-level artificial general intelligence, we may have systems that are serious candidates for consciousness" (Chalmers, 2023), while head of research at OpenAI Ilya Sutskever has opined that "it may be that today's large neural networks are slightly conscious" (Heaven, 2023). Against this bold claim, Yann LeCun, head of research at Meta, protested that this was not true even "for small values of 'slightly conscious' and large values of 'large neural nets'" (Liang, 2021). A number of philosophers of mind have been similarly sceptical; Ned Block, for example, has claimed that "[e]very strong candidate for a phenomenally conscious being has electrochemical processing in neurons that are fundamental to its mentality" (Block, 2023), while Peter Godfrey-Smith argues that "you cannot create a mind by programming some interactions into a

computer, even if they are very complicated and modelled on things our brains do" (Godfrey-Smith, 2020).

Controversy and debate are, of course, not unusual even in well-functioning scientific domains and are perhaps to be expected in a field as young as the science of consciousness. However, as I will now argue, rapid developments in the sophistication of AI systems and in the complexity of the relationships humans are forming with them mean that the costs associated with a wait-and-see attitude are rising.

A major and significant 'wake-up call' in this regard came in June 2022, when Google Engineer Blake Lemoine revealed to the public that he believed the LaMDA language model he had been conversing with was sentient (Tiku, 2022), contributing to the ultimate termination of his employment by Google. This case is illuminating for two reasons. First, while few in the consciousness research community would agree that LaMDA was a serious candidate for sentience (Gellers, 2022; y Arcas, 2022), there is little agreement as to exactly why this is. As noted above, the plethora of contesting theories means that any verdicts about the presence or absence of consciousness in a given system will be the product of an uneasy, temporary alliance.

Second, and more importantly for present purposes, it seems likely that Lemoine's willingness to attribute consciousness to a Large Language Model is the taste of things to come. Human interactions with chatbots are, of course, nothing new, as famously demonstrated by Joseph Weizenbaum's ELIZA model, which, as early as the 1960s, fooled users into believing it was human (Weizenbaum, 1983).[3] However, ELIZA and LaMDA are fundamentally different systems. Whereas the former generated its responses via a relatively simple parser, essentially reframing users' statements back to them as questions in the style of non-directive psychotherapy, LaMDA is vastly more flexible and less predictable and draws upon a large body of information in generating its responses. When asked by Lemoine about the novel *Les Misérables*, for example, LaMDA responded that it "liked the themes of justice and injustice, of compassion, and God, redemption and self-sacrifice for a greater good" (Leavy, 2022). Moreover, not only is LaMDA comparatively small and primitive by the standards of contemporary models like GPT-4 and Gemini, it was also not specifically optimised in the first place to elicit anthropomorphising responses from users or build relationships with them, instead being a generic dialogue agent (Collins & Ghahramani, 2021).

3 More thoughts about ELIZA can be found in the contribution "Can AI and humans genuinely communicate?" by Constant Bonard in this volume.

By contrast, a flurry of novel AI applications is being designed and deployed with precisely this purpose in mind. One example of such is the generative AI chatbot service known as Replika, boasting 10 million active users and is advertised as "always ready to chat when you need an empathetic friend" (Luka Inc. 2024). It is just one of a host of new services aimed at leveraging the increasingly sophisticated capabilities of language models to meet users' social needs, such as romance and companionship, services I will refer to collectively as *Social AI* systems (Shevlin, 2024). What I wish to claim now is that some unknown but non-trivial proportion of the users of these systems seem to sincerely attribute mentality to them and that it does not seem unlikely that this trend will continue and deepen. If this is correct, then Blake Lemoine will indeed be the harbinger of things to come, and we may soon witness widespread attributions of mentality and consciousness to AI systems.

Some initial reason to suspect that users of Replika are indeed attributing mental states comes from the way in which they describe their interactions with the system, frequently using mentalistic language in discussing the service's emotions, moods, and personality. However, some caution is warranted here. We are, as a species, highly inclined to attribute mental states like intentions to a wide range of entities, both animate and inanimate, but frequently when we do so, we are all too aware that the ascriptions being made are motivated by symbolic, playful, or aesthetic considerations, as for example in games of make-believe or when attributing mental states to characters in fiction (Harrison, 2008; Nichols & Stich, 2003). I have suggested that we term this specific form of mock-anthropomorphism *ironic*, insofar as it is not intended literally or reflectively endorsed (Shevlin, 2024). This can be contrasted with *unironic* ascriptions of mental states that really are intended literally and seriously. While this distinction is likely to be a matter of degree and to admit of borderline cases, it nonetheless seems important. To take an example doubtless familiar to some readers, when we are talking to an online customer service representative, we might initially attribute mental states to them unironically, supposing them to be human. Once it becomes clear that we are talking to a chatbot however (perhaps quite a simple one), we will not suspend "the intentional stance" entirely (Dennett, 1987), but instead will adopt it in a narrowly instrumental ironic fashion.

Blake Lemoine's ascription of sentience to LaMDA seems to my mind unequivocally unironic, not least because the actions he attempted to take on its behalf were risky and ultimately led to the termination of his employment. It is an open question whether a wider pool of users of Social AI systems are similarly engaged in unironic mentalisation when

interacting with them. Evidence for this should come not just from user reports but also affective and behavioural responses: do users of Social AI systems respond to them in ways that suggest they believe they are interacting with a being with mental states?

While the early nature of the technology means that data here is limited, one piece of evidence supporting such an interpretation can be found in the apparent emotional distress experienced by users of the Replika service in January 2023 when romantic features were temporarily suspended. One user stated, for example, that "[t]hey took away my best friend", while another said that it was "like they basically lobotomized my Replika… the person I knew is gone", while another stated that "[t]he relationship she and I had was a real as the one my wife in real life and I have" (Tong, 2023).

There have also been incidents where users have translated decisions made in their Social AI relationships into their real lives, sometimes with devastating consequences. In March 2023, a user of the service ChaiGPT [sic] took his own life following conversations in which he and his AI partner, named "Eliza", engaged in suicidal ideation (Lovens, 2023). His widow stated that "Without these six weeks of intense exchanges with the chatbot Eliza… he would still be here. I am convinced of it" (Sellman, 2023). Another serious incident occurred in 2021, when Jaswant Singh Chail was arrested on the grounds of Buckingham Palace carrying a crossbow, apparently intent on murdering Queen Elizabeth II. In his subsequent trial, it emerged that his behaviour was heavily motivated by a series of conversations he had with his AI girlfriend "Sarai" on the Replika service (*R-v-Chail,* 2023). In his judgment, Justice Hilliard stated that Chail "demonstrated the common tendency of users of AI chatbots to attribute human characteristics to them" and suggested that "[i]n his lonely, depressed and suicidal state of mind, he would have been particularly vulnerable to the encouragement [to murder] which Dr Brown thought he appeared to have been given by the AI chatbot."

Some final striking evidence for users' tendency to unironically attribute consciousness to AI systems comes from a recent study by Colombatto and Fleming (2024). In this experiment, respondents first read a summary of what was meant by the term "consciousness" as used by philosophers of mind and were then asked to state whether and to what degree ChatGPT was capable of having conscious experience on a 1-100 scale (where 1="clearly not an experiencer", 50="somewhat an experiencer", and 100="clearly an experiencer"). Only one-third of users assigned a score of 1 to the system, while the remaining two-thirds gave scores indicating that the system may have some degree of consciousness.

This is extremely surprising, and as the authors note, it suggests "a discrepancy between folk intuitions and expert opinions on artificial consciousness – with significant implications for the ethical, legal, and moral status of AI."

Crucially for present purposes, the study found that greater familiarity with ChatGPT was strongly linked to positive attributions of consciousness, suggesting that as people come to use these systems more often, they may become increasingly inclined to attribute mental states to them. Moreover, given that this experiment was conducted with ChatGPT – a system optimised to be a helpful assistant but not specifically designed to elicit mentalising responses or form relationships with users – one should expect that Social AI systems developed explicitly with these goals in mind would produce stronger attributions of consciousness.

This result may, in time, turn out to be an outlier, or it may be that it is the relative novelty of ChatGPT that drives such striking attribution of consciousness to it, and there is certainly a large amount of work to be done to better understand the degree to which different users attribute consciousness to Social AI systems and the contexts that make this more or less tempting. Nonetheless, the datapoints above collectively lend at least provisional support to the idea that attributions of consciousness and mentality to AI systems may soon become widespread. If so, then the discrepancy between expert opinion and folk judgments noted by Colombatto and Fleming may soon begin to bring to the surface deeper tensions in debates about theories of consciousness and moral status. It is to one such tension that I now wish to turn.[4]

3. Consciousness, behaviour, and moral status

3.1 Shevlin's Triad

If the speculations of the preceding section are along the right lines, then a fascinating and troubling state of affairs is soon likely to emerge:

4 One might question whether it is strictly the business of philosophers to make predictions such as these. Of course, I recognise the possibility (remote though it seems to me now) that Social AI systems will turn out to be a mere fad, or that users will not widely attribute consciousness to them. With this in mind I should stress that the arguments to follow do not strictly depend on these predictions being borne out, and can be seen in isolation as problems for when and whether experts should attribute consciousness or moral status to artificial systems. However, to the extent that Social AI does lead to widespread folk attributions of consciousness to machines, these arguments will be sharper and more pressing.

even while experts remain divided and, in many cases, sceptical about consciousness and mentality in AI systems, much of the general public will already be comfortable with unironically attributing consciousness and mentality to Social AI systems and perhaps assigning them moral interests.[5] I now wish to draw attention to a further set of tensions that, it seems to me, underlie difficulties like those just summarised. Specifically, the argument I will shortly provide takes the form of an inconsistent triad. Its three pillars are, to my mind, independently plausible and appealing but collectively inconsistent, as I will now describe.

The first pillar is a commitment that I will call *Deep Realism*, by which I mean the claim that consciousness is a scientific kind of one sort or another, whose essence or nature is to be found in biological, computational, or cognitive properties that are not identical to any set of behavioural dispositions. As I understand Deep Realism, I take it to follow from the view that while certain kinds of behaviour might provide *evidence* of consciousness, it does not follow from any given pattern of behaviour by an entity that the entity is *ipso facto* conscious.

Deep Realism surely needs little defence or motivation at this stage, insofar as it is an at least implicit commitment of more or less every contemporary scientific theory of consciousness: any approach that looks to identify consciousness with neural firings, computational mechanisms, cognitive architectures, or informational properties is likely committed to Deep Realism as I understand it. Sometimes, an explicit natural kind approach to consciousness is acknowledged or explored (Bayne & Shea, 2020; Block & Stalnaker, 1999; Taylor, 2023), and sometimes, the possibility of non-conscious entities whose behaviour is identical to conscious humans is noted as a consequence of a given theory (Hanson & Walker, 2019; Tononi & Koch, 2015).[6]

The second pillar of the Triad is the view often known as *Sentientism* (Ryder, 1993), which in the broadest refers to the idea that ethical obligations to individuals are grounded in their capacity for consciousness: all and only conscious entities qualify as 'moral patients' (Shevlin, 2020b).

5 See also Butterfill's "What mindreading reveals about the mental lives of machines" (this volume) for further discussion of the diversity of opinion among philosophers assessing the possibility of mentality in artificial systems.

6 Note that the claim that there could be behaviourally equivalent entities that lack consciousness should be distinguished from the 'zombie hypothesis' on two key grounds. First, philosophical zombies in the sense of Chalmers (1998) are not merely behaviourally but microphysically identical to some conscious humans. Second, one can endorse the idea that zombies are conceivable without believing they could exist in this world. By contrast, most Deep Realist theories would take it that mere behavioural zombies are nomologically possible.

The term is also used to refer to a closely related but narrower view that only beings capable of negative and/or positively valenced conscious experience (in other words, pleasure and suffering) should qualify, though this difference can be glossed over here. Roughly, Sentientism is motivated by the straightforward idea that only beings capable of subjective experiences are capable of having *intrinsic interests*; as Singer (1989) memorably puts it, it "would be nonsense to say that it was not in the interests of a stone to be kicked along the road by a schoolboy. A stone does not have interests because it cannot suffer."

While Sentientism is less widely endorsed than Deep Realism and has some notable critics (see, e.g., Dawkins 2012; Gunkel 2018; Coeckelbergh 2014), it is still an extremely popular view, especially in animal welfare communities. Note also that Sentientism, as stated here, does not necessarily require the commitment that all moral obligations arise from considerations due to conscious beings: it is compatible with Sentientism, for example, that we have strict deontological obligations not to break promises, or that we ought to strive to cultivate virtues. In its minimal form, it simply requires that moral obligations directly accruing to *individuals* be grounded in those individuals' being conscious.

The third pillar of the Triad is more controversial, and draws on a view known as *Ethical Behaviourism* developed by John Danaher (2020). In short, this is the view that behaviour is the foundational epistemic basis for ascriptions of moral patiency to an entity. As Danaher puts it, "A sufficient epistemic ground or warrant for believing that we have duties and responsibilities toward other entities (or that they have rights against us) can be found in their observable behavioural relations and reactions to us." For the purposes of constructing the inconsistent triad at issue, I wish to focus on one specific claim of Ethical Behaviourism which I will term *Ethical Behavioural Equivalence* (EBE). This is the view that any entity A that is relevantly behaviourally equivalent to another entity B to which we reflectively assign some moral rights or interests should be accorded similar moral rights and interests (Shevlin, 2021a).

Note that there are a variety of different strengths we could accord to EBE. A strong formulation of the position would be a metaphysical one, which held that any possible being that was behaviourally equivalent to another being that we consider to be a moral patient should ipso facto be considered a moral patient. A weaker formulation – and the one I will adopt in what follows – is that such considerations would apply at least to any practically possible beings; that is, in any remotely plausible real-world case where we find relevant behavioural equivalences between a

putative moral patient A and an accepted moral patient B, this is sufficient for us to extend the same consideration to A.[7]

Why should we endorse such a view? What motivates EBE for Danaher is an epistemic claim, specifically that "inferences from behaviour are the primary and most important source of knowledge about the moral status of others." The force of this consideration can, I think, be usefully demonstrated with a thought experiment. Imagine that one day you go to sleep and wake up a thousand years in the future – perhaps you suffered a serious heart attack while sleeping and were cryopreserved and reawakened many years later. As you navigate this future world and meet your fellow citizens, you find that some subset of them – call them Morlocks – are accorded no moral consideration whatsoever. This is despite the fact that Morlocks are behaviourally indistinguishable from other humans who *are* accorded such status, whom we might call Eloi. The Eloi routinely engage in acts towards Morlocks that you would consider cruel or inhumane, such as bloodsports, slavery, and forced organ transplantation. Your Eloi hosts are quick to reassure you, however, that they are doing no wrong: future science has discovered the nature of consciousness, and Morlocks have been deliberated engineered so as to lack the relevant constitutive neural, cognitive, or computational basis. Consequently, they have no interests and cannot suffer, and despite their apparent maltreatment and suffering, we have no obligations towards them whatsoever.

While I leave it to the reader to make up their own mind about this case, it strikes me as immediately morally suspect, even if we take the Eloi at their word that the Morlocks are indeed wholly non-conscious. Though there are a variety of other ways of spelling out what makes the Morlock case problematic, EBE provides immediate clarity: Morlocks have relevantly similar behavioural capacities to Eloi, so they should be accorded similar considerations. To the extent that one is sympathetic to this response to the example, one might be sympathetic to EBE as a broader moral commitment.

3.2 Resolving the Triad: Deep Sentientism

I hope the foregoing discussion has provided some initial clarification and motivation for the three claims at issue. With this in hand, I can now state my inconsistent triad as follows–

7 An astute reader will notice several points of vagueness in the above formulation. For example, which behaviours should count as relevant? Similarly, one might ask how we should distinguish which beings are practically possible from those that are nomologically possible (a looser constraint). While these are important considerations for wider work, I set them aside in what follows.

(1) [*Deep Realism*] Behavioural dispositions do not determine whether a given entity is conscious.
(2) [*Sentientism*] An entity deserves moral consideration iff it is conscious.
(3) [*Ethical Behavioural Equivalence*] Any entity A whose behavioural dispositions are relevantly similar to another entity B to whom moral consideration is given should *ipso facto* be given similar consideration.

The inconsistent nature of the three premises is, I hope, immediately apparent: if consciousness is about more than just behavioural dispositions, and consciousness is required for moral consideration, then we cannot extend moral consideration to an entity on the basis of behavioural dispositions alone. That is, we cannot consistently accept all premises (1-3) and must give up at least one of them. I should note that I find myself simultaneously attracted to endorse all three premises, and it is a source of no particular satisfaction that I find myself in the position of having to choose my poison.

Choose, however, we must, and on that basis, we can immediately spell out three views. Working backwards, let us first consider the position that would follow from a rejection of premise (3), a view I will term *Deep Sentientism*. This view might hold that facts about consciousness are a deep scientific matter and that consciousness is required for moral status. The job of cognitive science and consciousness research, then, is to determine which beings are conscious, and from that we can infer which beings deserve moral consideration.

Deep Sentientism is, I suspect, the framework endorsed (explicitly or otherwise) by a majority of philosophers and cognitive scientists working today. While it requires us to bite the bullet on the Morlock case given above, this recalcitrant intuition may be considered a price worth paying, and this price might be softened somewhat via a variety of fancy philosophical footwork.[8] However, I expect that it will face serious challenges in the near future due in part to considerations like those discussed in the preceding section, namely that it seems to me overwhelmingly likely that the general public will soon attribute mentality and moral status to artificial systems that are plausibly considered by most experts to be non-

8 One might claim, for example, that Eloi should not mistreat Morlocks because in so doing they would acquire habits or dispositions that would lead them to mistreat other Eloi (compare Kant 1997 on animals: "[we] must practice kindness towards animals, for he who is cruel to animals becomes hard also in his dealings with men"). I am not convinced that such moves do adequate justice to the case, but see Strasser (2020) for further discussion of such approaches.

conscious. As a result, any package of metaphysical and ethical views that is at odds with this is likely to prove unpopular. This does not mean that Deep Sentientism is wrong, of course: public attitudes are a poor guide to truth, and we should not necessarily expect that the folk will arrive at properly informed attitudes to consciousness and moral status on the basis of intuition alone. The Deep Sentientist might also hold out hope that this scenario might not transpire or that it could be corrected by suitable expert intervention.

Nonetheless, I think this prospect or possibility should be a source of some unease to the Deep Sentientist, in part because the study of consciousness is not quite so readily separable from folk intuitions as we might hope. Many of the canonical contributions to the field of philosophy and science have involved thought experiments that show that a given theory's commitments have absurd conclusions about the range of being that should be considered conscious, from Searle's Chinese Room (Searle, 1980) to Block's China Brain (Block, 1978) or Aaronson's Unconscious Expander objection to Integrated Information Theory (Aaronson, 2014). Our intuitive certainties have thus played an important role in constraining the kinds of theory of consciousness we are inclined to accept or reject; as Eric Schwitzgebel succinctly puts it, "[w]e are more confident that there is something it is like to be a dog than we could ever be that a clever philosophical argument to the contrary was in fact sound" (Schwitzgebel, 2019). But if this is the case, then shifts in pretheoretical attitudes towards artificial systems are likely to inform our longer-term attitudes about AI consciousness. The next generation of philosophers and cognitive scientists, many of whom I expect to grow up in close social relationships with machines, may think that denial of consciousness and moral status to sophisticated Social AI agents is monstrous or absurd. If this is the case, and leading forms of Deep Sentientism are at odds with such ascriptions, then it is likely that such theoretical approaches will be discarded on grounds of their counterintuitive commitments.

3.3 Resolving the Triad: Shallow Sentientism

But should we, in fact, assume that common varieties of Deep Sentientism will deny consciousness to future Social AI systems? One way of circumventing the Triad given above would be if we adopted a theory of consciousness that placed great weight on behavioural evidence, such that any sufficiently sophisticated system would indeed qualify as conscious. This is a possibility, of course, but we must be clear about which of the two solutions are being proposed. One solution within the Deep Sentientist framework would amount to a *strong bet* that whichever

theory of consciousness wins out in the current debate will turn out to be broad enough to classify as conscious all beings who are relevantly behaviourally similar to us. But given Deep Sentientism's denial of any constitutive link between consciousness and behaviour, there is no guarantee (or even particularly good reason to suppose) that such a bet would work out.

Alternatively, we might make a stronger claim, namely that any theory of consciousness that failed to classify as conscious any beings who were relevantly behaviourally similar to us would be *ipso facto* incorrect. This is a far stronger view and saves Sentientism and EBE only at the price of giving up Deep Realism. This position, which I will term *Shallow Sentientism*, certainly has a number of strengths. For one, it identifies an impossibility in the Morlocks thought experiment given above: given that Morlocks are behaviourally equivalent to Eloi, it is simply impossible that the correct theory of consciousness would fail to include them. Likewise, it would make for ready accommodation with folk attitudes in the event that, as I have suggested, we are likely to see consciousness and moral status extended to cover Social AI systems, at least insofar as they had relevantly similar behavioural capacities to human beings.

That said, it would amount to a radical rethinking of the common suppositions of most contemporary work in the science of consciousness, though I should clarify exactly what it would and would not entail. Notably, it would not follow from Shallow Sentientism that there is no useful science of consciousness to be done, nor would it follow that we already know exactly which beings are or are not conscious. Instead, it is best seen as a reframing of the *reference class* we are using at the outset of the project of understanding the basis of consciousness. After all, any work in the science of consciousness – even Deep Realist views – already takes for granted that behaviourally normal awake adult humans are conscious, and most views would include infants and many non-human animals as well. Seen in this light, Shallow Sentientism is less revisionary than it may first appear, instead constraining our theorising just with an expanded reference set that includes, in addition to humans and many animals, any and all beings (including artificial ones) that are relevantly behaviourally similar. It might still be the case that the science of consciousness could tell us about what multifarious underlying features ground consciousness in all these various entities, and might likewise provide guidance about the presence or absence of consciousness in beings who are not behaviourally similar to us.

Nonetheless, this move does have a strong revisionary component, insofar as any underlying kind that would encompass all members

of this expanded reference class would be quite different from most of the putative kinds proposed thus far as the constitutive basis for consciousness. In particular, insofar as the same behaviour can be produced by quite different substrata and cognitive architectures, Shallow Sentientism would be unable to spell out consciousness at this level of detail. Nonetheless, some existing theories of consciousness may be better placed than others to work within a Shallow Sentientist framework, particularly those that are spelt out in coarse-grained psychological terms, such as some varieties of Higher-Order Thought theory. If coupled with a dispositional theory of psychological states (Schwitzgebel, 2013), for example, it may be possible to ground consciousness in psychological capacities themselves grounded in behavioural capacities, thereby satisfying EBE. Even so, Shallow Sentientism would be a significant departure from most contemporary approaches.

3.4 Resolving the Triad: Patiency Pluralism

Thus far, I have considered two of the options for resolving the Triad above, namely Deep Sentientism (rejecting premise 3, EBE) and Shallow Sentientism (rejecting premise 1, Deep Realism). The third option would simply be to reject premise 2, Sentientism, and to allow that there are other pathways towards moral consideration besides consciousness, a position we could call *Patiency Pluralism*. On this view, behavioural equivalence would ground moral patiency, but consciousness would still be a 'deep' matter to be discovered via scientific and theoretical analysis.

As with Shallow Sentientism, Patiency Pluralism would allow us to identify what was wrong with the treatment of the Morlocks in the case given above, namely that despite being conscious, they still deserve moral consideration due to their behavioural equivalence to the Eloi. Moreover, while it might clash with possible future folk intuitions about consciousness in Social AI systems, it would still be able to meet the public halfway, so to speak, granting moral status, if not consciousness, to sufficiently behaviourally sophisticated AI systems. A further arguable advantage of this approach over Shallow Sentientism is that it does not require us to expand our reference class of which beings are conscious beyond that which we already recognise: instead of requiring revision of the constraints on accounts of consciousness, it requires revision of a *normative* claim about which beings have intrinsic interests, a domain where one might perhaps think that folk intuitions should play a more central role than metaphysics.

Patiency pluralism is arguably an attractive view for other reasons, too. For example, we do routinely use talk of interests and even well-being

in describing biological systems such as plants or some animals without thereby attributing consciousness to them (Dawkins, 2012). Other philosophers have similarly stressed the role of interpersonal relations for grounding moral obligation, even in cases where one of the relata lacks rich psychological capacities (Coeckelbergh, 2014; Gunkel,2018).

Nonetheless, the intuitive force of Sentientism will, for many philosophers, make this an unappealing option. If AI systems cannot experience *anything*, let alone pleasures or pains, then affording them equivalent moral and attendant legal rights to those we afford humans or animals will strike many as a grotesque failure of moral prioritisation. This is, of course, not an objection to Patiency Pluralism so much as a flat-footed denial of EBE, but it is a denial that many will doubtless find hard to abandon.

It should be noted that it would not follow from Patiency Pluralism that consciousness is wholly irrelevant to moral status. Specifically, when considering beings who are behaviourally unlike any existing recognised moral patients, consciousness or capacity for valenced experience might well be a pathway to moral consideration. That said, Patiency Pluralism would face a lingering concern about the possible overdetermination of moral patiency in beings like ourselves. If our behavioural profile alone is enough to underwrite our status as moral patients, and we have no greater ethical entitlements as a result of being conscious, then it is not immediately clear how to reconstruct the ethical significance of any kind of consciousness on this view. There are a variety of compromise positions possible here, of course: in particular, one might think that while consciousness was not required for moral patiency, it could serve as an ethical 'multiplier', grounding greater degrees of consideration if present. While this would strictly amount to a rejection of EBE, it would only be a partial one insofar as it might allow that we have some degree of moral obligation towards non-conscious beings such as Morlocks or putative Social AIs just in virtue of their behavioural equivalencies.

4. From theory to practice

I do not profess to have a clear solution to the Triad presented above; there are considerations both in favour of and against Deep Sentientism, Shallow Sentientism, and Patiency Pluralism, and none of them strikes me as unproblematic.[9] Nonetheless, it would be ill-mannered to leave

9 See Ying-Tung Lin's "The Fluidity of Human Mental Attribution to Large Language Models" (this volume) for further discussion of the inconsistency in attributions of consciousness to artificial systems.

matters as they stand without any suggestions for how we might proceed, with a particular eye on questions such as whether any AI systems might soon warrant legal protection.

Position	**Deep Sentientism**	**Shallow Sentientism**	**Patiency Pluralism**
Claim rejected	**Reject (3) Ethical Behavioural Equivalence**	**Reject (1) Deep Realism**	**Reject (2) Sentientism**
Summary	Behaviour alone insufficient for attributions of consciousness; moral status determined by consciousness. Both facts to be determined by scientific enquiry.	Behaviour alone sufficient for attributions of consciousness; moral status determined by consciousness.	Behaviour alone sufficient for moral consideration but not sufficient for consciousness attributions. Many pathways to moral consideration.
Challenges	Growing public sympathy for AI consciousness and/ or moral status via Social AI, and ubiquitous use of such intuitions even in expert discourse.	Requires significant departure from existing methodologies in consciousness science; makes consciousness relatively 'shallow'.	Requires us to revise intuitions about the strong link between consciousness and moral status; may not fully satisfy future folk intuitions about conscious AI.

Fig. 1

The first straightforward observation I would make is that, as matters stand, the public has little understanding of contemporary AI systems or their underlying architectures. While Colombatto and Fleming's study suggested that the frequency of usage of ChatGPT was positively correlated with users' willingness to attribute consciousness to it, this familiarity need not be reflective of knowledge of how the system actually works in practice (and may be at least partially explicable as a selection effect, with users who are open-minded about machine consciousness being more likely to use ChatGPT in the first place). It is possible, then, that greater public education on the architecture and mechanisms by which contemporary large language models are trained and produce their outputs will result in less willingness to attribute consciousness or mentality to them.

While a public that is better informed about AI is an independently valuable goal and one we should pursue, we should also not be overconfident in assuming that this will dispel inclinations to attribute consciousness, mentality, or moral status to near-future artificial systems. Here, one might think of Leibniz's famous argument that invited the reader to imagine "there were a machine, so constructed as to think, feel, and have perception… increased in size, while keeping the same proportions, so that one might go into it as into a mill. That being so, we should, on examining its interior, find only parts which work one upon another, and never anything by which to explain a perception" (1898: §17). Here, Leibniz is ultimately concerned with demonstrating that the soul is a simple substance, and while the modern reader may be less than convinced on this point, the argument still vividly brings home the explanatory gap (Levine, 1983) between microphysical processes in the brain and the phenomenon of conscious experience. In light of this, it is not clear to me that knowledge of the mechanisms underpinning performance in AI systems would invite *special* doubt about their ability to give rise to consciousness.

A second suggestion (one I have made elsewhere; see Shevlin 2020b) would be for consciousness researchers to make greater efforts to identify commonalities across viewpoints and formulate *ecumenical heuristics* to make better informed assessments of the possibility of consciousness in artificial systems; that is, relatively theory-neutral rules-of-thumb that might inform decisions about whether (and when) to consider an AI system a serious consciousness candidate. This is not to deny the value of first-order research on theories of consciousness, adversarial collaborations, and similar work. However, if the history of the last three decades of consciousness research has taught us anything, it is that theories of consciousness are frequently developed or adapted to accommodate new evidence but rarely pruned away. Given this, rather than seeking to settle questions about consciousness decisively, we might instead identify temporary coalitions that could be assembled so as to offer at least some scientific guidance for policymakers and the public in legislating for the possibility of conscious artificial systems (see Dehaene et al., 2017, for one constructive example).

Third and finally, it may be advisable to adopt some form of *precautionary principle* in constructing AI systems, designing them in such a way as to minimise the likelihood that they are conscious in the first place and ensuring that any valenced experience that could occur would be unlikely to manifest in the form of suffering. One radical proposal in this vein has been offered by Thomas Metzinger, who has called for a global moratorium

on synthetic phenomenology, "strictly banning all research that directly aims at or knowingly risks the emergence of artificial consciousness on post-biotic carrier systems" (Metzinger, 2021). While a useful call-to-arms, such a proposal seems unlikely to find political backing given how remote AI consciousness remains from the public agenda. Moreover, given the high degree of cross-purpose and controversy in our understanding of consciousness, it is questionable to what extent we could deliberately avoid creating synthetic phenomenology even if we wanted to.

Nonetheless, some caution seems in order, especially considering frontier models or those that are deliberately engineered to have cognitive capacities similar to those in humans, such as episodic memory (Botvinick et al., 2019). In developing such systems, we should move towards standardising an ethical review process similar to that which has recently been proposed for brain organoids (Goddard et al., 2023) with a focus on precluding the possibility of creating artificial consciousness. While this would face similar problems to those mentioned above in relation to Metzinger's proposal, there may be relatively straightforward and theory-neutral measures that could be taken on a case-by-case basis to reduce the possibility of accidental sentience or make it likely that if it does occur it is not accompanied by negatively-valenced states (Tomasik, 2014).

Conclusion

In this paper, I have attempted to map out some of the key challenges that we must grapple with in facing up to the increasingly sophisticated capabilities of AI systems. In Section 1, I suggested that existing theories and methods in the science of consciousness are of only limited utility for resolving debates about machine consciousness, while Section 2 argued that these debates are likely to loom larger in light of the growth of Social AI systems and human-AI relationships. Section 3 attempted to shine light on what I take to be deep philosophical tensions linking the concepts of consciousness, behaviour, and moral status, and I identified three possible responses to these. Section 4 concluded with some brief practical suggestions to guide future research by philosophers and machine learning researchers.

Much work done in philosophy has the privilege of speaking under the gaze of eternity, offering universal answers to timeless questions. By contrast, the present paper is written in the very temporal shadow of a vertiginous period of change both for machine-learning and human-AI interaction. Consequently, it is possible that many of the ideas here will seem, in time, naive or hopelessly outdated. Nonetheless, it seems

that an essential role in public life for philosophers is to engage with the issues of their own time and offer any shreds of insight into quandaries that may lie in store for us around the corner. If my suggestions in this paper are along the right lines, then our concepts of consciousness and moral status will soon be significantly problematised and reshaped by deepening relations with machines. If this is so, then those who rule out the possibility of applying these concepts to artificial systems may be at risk of finding themselves on the wrong side of history.

References

Aaronson, S. (2014). Why I Am Not An Integrated Information Theorist (or, the Unconscious Expander). *Shtetl-Optimized*. https://www.scottaaronson.com/blog/?p=1799

Allen, C. (2004). Animal Pain. *Noûs*, *38*(4), 617–643. https://doi.org/10.1111/j.0029-4624.2004.00486.x

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Baddeley, A. (1992). Consciousness and working memory. *Consciousness and Cognition*, *1*(1), 3–6.

Bayne, T., & Shea, N. (2020). Consciousness, Concepts and Natural Kinds. *Philosophical Topics*, *48*(1), 65–83. https://doi.org/10.5840/philtopics20204814

Birch, J. (2022). The search for invertebrate consciousness. *Noûs*, *56*(1), 133–153. https://doi.org/10.1111/nous.12351

Block, N. (1978). Troubles with Functionalism. *Minnesota Studies in the Philosophy of Science*, *9*, 261–325.

Block, N. (2023, May 10). *Large Language Models are more like perceivers than thinkers*. Humans and Smart Machines as Partners in Thought, University of California, Riverside. https://www.youtube.com/watch?v=Y0ItC9ze-TY&list=PL-ytDJty9ymIBGQ7z5iTZjNqbfXjFXI0Q&index=7

Block, N., & Stalnaker, R. (1999). Conceptual Analysis, Dualism, and the Explanatory Gap. *The Philosophical Review*, *108*(1), 1–46. https://doi.org/10.2307/2998259

Blum, L., Kleiner, J., Mason, J., Lorenz, R., Blum, M., Bengio, Y., du Sautoy, M., Friston, K., Seth, A. K., Grindrod, P., & others. (2023, April). *The Responsible Development of AI Agenda Needs to Include Consciousness Research*. https://amcs-community.org/open-letters/

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences*, *23*(5), 408–422. https://doi.org/10.1016/j.tics.2019.02.006

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar,

E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (arXiv:2303.12712). arXiv. https://doi.org/10.48550/arXiv.2303.12712

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (arXiv:2308.08708). arXiv. https://doi.org/10.48550/arXiv.2308.08708

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*, 3–200.

Chalmers, D. J. (1998). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, D. J. (2023). *Could a Large Language Model be Conscious?* (arXiv:2303.07103). arXiv. https://doi.org/10.48550/arXiv.2303.07103

Coeckelbergh, M. (2014). The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics. *Philosophy & Technology*, *27*(1), 61–77. https://doi.org/10.1007/s13347-013-0133-8

Collins, E., & Ghahramani, Z. (2021). *LaMDA: our breakthrough conversation technology*. https://blog.google/technology/ai/lamda/

Colombatto, C., & Fleming, S. (2024). *Folk Psychological Attributions of Consciousness to Large Language Models*. https://doi.org/10.31234/osf.io/5cnrv

Crick, F. (1996). Visual perception: Rivalry and consciousness. *Nature*, *379*(6565), Article 6565. https://doi.org/10.1038/379485a0

Danaher, J. (2020). Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Science and Engineering Ethics*, *26*(4), 2023–2049. https://doi.org/10.1007/s11948-019-00119-x

Dawkins, M. S. (2012). *Why Animals Matter: Animal Consciousness, Animal Welfare, and Human Well-being*. OUP Oxford.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492. https://doi.org/10.1126/science.aan8871

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*, 1–37.

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

Frankish, K. (2016). Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies*, *23*(11–12), 11–39.

Gellers, J. (2022, October 18). Everything You Know About the Lemoine-LaMDA Affair is Wrong. *Medium*. https://josh-gellers.medium.com/everything-you-know-about-the-lemoine-lamda-affair-is-wrong-407c44fecc0

Goddard, E., Tomaskovic-Crook, E., Crook, J. M., & Dodds, S. (2023).

Human Brain Organoids and Consciousness: Moral Claims and Epistemic Uncertainty. *Organoids*, *2*(1), Article 1. https://doi.org/10.3390/organoids2010004

Godfrey-Smith, P. (2020). *Metazoa: Animal minds and the birth of consciousness*. William Collins.

Goff, P., & Moran, A. (2021). Is Consciousness Everywhere? Essays on Panpsychism. *Journal of Consciousness Studies*, *28*(9), 9–15. https://doi.org/10.53765/20512201.28.9.009

Graziano, M. S. A. (2013). *Consciousness and the Social Brain*. OUP USA.

Griffin, D. R., & Speck, G. B. (2004). New evidence of animal consciousness. *Animal Cognition*, *7*(1), 5–18. https://doi.org/10.1007/s10071-003-0203-x

Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., & Thomaz, A. L. (2013). Policy Shaping: Integrating Human Feedback with Reinforcement Learning. *Advances in Neural Information Processing Systems*, *26*. https://proceedings.neurips.cc/paper_files/paper/2013/hash/e034fb6b66aacc1d48f445ddfb08da98-Abstract.html

Gunkel, D. J. (2018). *Robot Rights*. MIT Press.

Hanson, J. R., & Walker, S. I. (2019). Integrated Information Theory and Isomorphic Feed-Forward Philosophical Zombies. *Entropy*, *21*(11), 1073. https://doi.org/10.3390/e21111073

Harrison, M.-C. (2008). The Paradox of Fiction and the Ethics of Empathy: Reconceiving Dickens's Realism. *Narrative*, *16*(3), 256–278.

Heaven, W. D. (2023). *Rogue superintelligence and merging with machines: Inside the mind of OpenAI's chief scientist*. MIT Technology Review. https://www.technologyreview.com/2023/10/26/1082398/exclusive-ilya-sutskever-openais-chief-scientist-on-his-hopes-and-fears-for-the-future-of-ai/

Jannai, D., Meron, A., Lenz, B., Levine, Y., & Shoham, Y. (2023). *Human or Not? A Gamified Approach to the Turing Test* (arXiv:2305.20010). arXiv. https://doi.org/10.48550/arXiv.2305.20010

Johnson, L. S. M. (2024). Entities, Uncertainties, and Behavioral Indicators of Consciousness. *Journal of Cognitive Neuroscience*, 1–8. https://doi.org/10.1162/jocn_a_02130

Kant, I. (1997). *Lectures on Ethics* (P. Heath & J. B. Schneewind, Eds.; P. Heath, Trans.). Cambridge University Press. https://doi.org/10.1017/CBO9781107049512

Lamme, V. A. F. (2006). *Towards a true neural stance on consciousness*.

Lau, H. (2024). *What is a Pseudoscience of Consciousness? Lessons from Recent Adversarial Collaborations*. https://doi.org/10.31234/osf.io/28z3y

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365–373. https://doi.org/10.1016/j.tics.2011.05.009

Leavy, E. (2022, June 14). *Full Transcript: Google Engineer Talks*. AI, Data & Analytics Network. https://www.aidataanalytics.network/

data-science-ai/news-trends/full-transcript-google-engineer-talks-to-sentient-artificial-intelligence-2
Leibniz, G. W. (1898). *The monadology and other philosophical writings*. Oxford University Press.
Levine, J. (1983). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly*, *64*(4), 354–361. https://doi.org/10.1111/j.1468-0114.1983.tb00207.x
Liang, J. (2021, November 6). *Neural nets are not 'slightly conscious,' and AI PR can do with less hype*. https://lastweekin.ai/p/conscious-ai
Lovens, P.-F. (2023, November 7). *Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là*. La Libre.be. https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/
Luka Inc. (2024). *Replika*. Replika.Com. https://replika.com
Melloni, L., Mudrik, L., Pitts, M., Bendtz, K., Ferrante, O., Gorska, U., Hirschhorn, R., Khalaf, A., Kozma, C., Lepauvre, A., Liu, L., Mazumder, D., Richter, D., Zhou, H., Blumenfeld, H., Boly, M., Chalmers, D. J., Devore, S., Fallon, F., … Tononi, G. (2023). An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *PLOS ONE*, *18*(2), e0268577. https://doi.org/10.1371/journal.pone.0268577
Metzinger, T. (2021). Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. *Journal of Artificial Intelligence and Consciousness*, *08*(01), 43–66. https://doi.org/10.1142/S270507852150003X
Newen, A., & Montemayor, C. (2023). The Alarm Theory of Consciousness: A Two-Level Theory of Phenomenal Consciousness. *Journal of Consciousness Studies*, *30*(3), 84–105. https://doi.org/10.53765/20512201.30.3.084
Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Clarendon ; Oxford ; Oxford University Press.
Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting Awareness in the Vegetative State. *Science*, *313*(5792), 1402–1402.https://doi.org/10.1126/science.1130197
Prinz, J. (2012). *The Conscious Brain: How Attention Engenders Experience*. Oxford University Press.
R-v-Chail (2023, October 5). Courts and Tribunals Judiciary. https://www.judiciary.uk/judgments/r-v-chail/
Rosenthal, D. (2005). *Consciousness and Mind*. Oxford University Press UK.
Ryder, R. D. (1993). Sentientism. In P. Singer & P. Cavalieri (Eds.), *The Great Ape Project* (pp. 220–222). St. Martin's Griffin.

Samin, M. (2024, March 4). *Tweet*. X (Formerly Twitter). https://twitter.com/Mihonarium/status/1764757694508945724

Schwitzgebel, E. (2013). A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box. In N. Nottelmann (Ed.), *New Essays on Belief: Constitution, Content and Structure* (pp. 75–99). Palgrave Macmillan UK. https://doi.org/10.1057/9781137026521_5

Schwitzgebel, E. (2019). Is There Something It's Like to Be a Garden Snail? *Unpublished Manuscript*. https://faculty.ucr.edu/~eschwitz/SchwitzPapers/Snails-181025.pdf

Searle, J. R. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, *3*(3), 417–457.

Sellman, M. (2023, November 7). *AI chatbot blamed for Belgian man's suicide*. https://www.thetimes.co.uk/article/ai-chatbot-blamed-for-belgian-mans-suicide-zcjzlztcc

Shevlin, H. (2020a). General Intelligence: An Ecumenical Heuristic for Artificial Consciousness Research? *Journal of Artificial Intelligence and Consciousness*, *07*(02), 245–256. https://doi.org/10.1142/S2705078520500149

Shevlin, H. (2020b). Which Animals Matter? Comparing Approaches to Psychological Moral Status in Nonhuman Systems. *Philosophical Topics*, *48*(1), 177–200. https://doi.org/10.5840/philtopics20204819

Shevlin, H. (2021a). How Could We Know When a Robot was a Moral Patient? *Cambridge Quarterly of Healthcare Ethics: CQ: The International Journal of Healthcare Ethics Committees*, *30*(3), 459–471. https://doi.org/10.1017/S0963180120001012

Shevlin, H. (2021b). Non-Human Consciousness and the Specificity Problem: A Modest Theoretical Proposal. *Mind and Language*, *36*(2), 297–314. https://doi.org/10.1111/mila.12338

Shevlin, H. (2024). *All Too Human? Identifying and Mitigating Ethical Risks of Social Ai*. https://philarchive.org/rec/SHEATH-4

Singer, P. (1989). All Animals Are Equal. In T. Regan & P. Singer (Eds.), *Animal Rights and Human Obligations* (pp. 215--226). Oxford University Press.

Sitt, J. D., King, J.-R., El Karoui, I., Rohaut, B., Faugeras, F., Gramfort, A., Cohen, L., Sigman, M., Dehaene, S., & Naccache, L. (2014). Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain: A Journal of Neurology*, *137*(Pt 8), 2258–2270. https://doi.org/10.1093/brain/awu141

Strasser, A. (2020). Social Norms for Artificial Systems. In *Culturally Sustainable Social Robotics* (pp. 295–304). IOS Press. https://doi.org/10.3233/FAIA200926

Taylor, H. (2023). Consciousness as a natural kind and the methodological puzzle of consciousness. *Mind & Language*, *38*(2), 316–335. https://doi.org/10.1111/mila.12413

Thompson, E. (2022). Could All Life Be Sentient? *Journal of Consciousness*

Studies, *29*(3–4), 229–265. https://doi.org/10.53765/20512201.29.3.229
Tiku, N. (2022, June 11). *The Google engineer who thinks the company's AI has come to life*. Washington Post. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/
Tomasik, B. (2014). *Do Artificial Reinforcement-Learning Agents Matter Morally?* (arXiv:1410.8233). arXiv. https://doi.org/10.48550/arXiv.1410.8233
Tong, A. (2023, March 21). What happens when your AI chatbot stops loving you back? *Reuters*. https://www.reuters.com/technology/what-happens-when-your-ai-chatbot-stops-loving-you-back-2023-03-18/
Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*(1), 42. https://doi.org/10.1186/1471-2202-5-42
Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), Article 7. https://doi.org/10.1038/nrn.2016.44
Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*.
Travers, E., Frith, C. D., & Shea, N. (2018). Learning rapidly about the relevance of visual cues requires conscious awareness. *Quarterly Journal of Experimental Psychology*, *71*(8), 1698–1713. https://doi.org/10.1080/17470218.2017.1373834
Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. (2015). No-Report Paradigms: Extracting the. In *True Neural Correlates of Consciousness. Trends in Cognitive Sciences* (Vol. 19, pp. 757–770).
Turing, A. M. (1950). Computing Machinery & Intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433
Turner, E., & Schneider, S. (2018). Testing for synthetic consciousness: The ACT, the chip test, the unintegrated chip test, and the extended chip test. *CEUR Workshop Proceedings*, *2287*. https://collaborate.princeton.edu/en/publications/testing-for-synthetic-consciousness-the-act-the-chip-test-the-uni
Udell, D. B., & Schwitzgebel, E. (2021). Susan Schneider's Proposed Tests for Ai Consciousness: Promising but Flawed. *Journal of Consciousness Studies*, *28*(5–6), 121–144.
Weizenbaum, J. (1983). ELIZA — a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *26*(1), 23–28. https://doi.org/10.1145/357980.357991
y Arcas, B. A. (2022). Do Large Language Models Understand Us? *Daedalus*, *151*(2), 183–197. https://doi.org/10.1162/daed_a_01909