Murray Shanahan

Satori Before Singularity

Abstract: *According to the singularity hypothesis, rapid and accelerating technological progress will in due course lead to the creation of a human-level artificial intelligence capable of designing a successor artificial intelligence of significantly greater cognitive prowess, and this will inaugurate a series of increasingly super-intelligent machines. But how much sense can we make of the idea of a being whose cognitive architecture is qualitatively superior to our own? This article argues that one fundamental limitation of human cognitive architecture is an inbuilt commitment to a metaphysical division between subject and object, a commitment that could be overcome in an artificial intelligence lacking our biological heritage.*

1. Introduction

'If a lion could talk we would not understand him', Wittgenstein famously remarks in the *Philosophical Investigations* (Wittgenstein, 1958, p. 223). His point is that understanding is predicated on a shared form of life, where 'form of life' encompasses everything that goes to make up the world for an organism, including its biology, its values, its culture, and so on. In this sense, though living on the same planet as ourselves, Wittgenstein's lion inhabits a different world. How much more inscrutable to us, then, would be an AI+, the imagined product of a human-level artificial intelligence that engineered (or morphed into) a successor of significantly greater cognitive sophistication?[1]

Correspondence:
Email: m.shanahan@imperial.ac.uk

[1] Good (1965); Moravec (1988); Vinge (1993); Kurzweil (2005); Chalmers (2010). Chalmers introduces the term AI+ to denote 'artificial intelligence of greater than human level (that is, more intelligent than the most intelligent human)'. The usage here makes less appeal to a notional scale of intelligence. To qualify as having significantly greater cognitive sophistication than a human being, the AI+ should not be attainable simply by making human-level AI faster, larger (in any relevant sense), or more numerous.

Journal of Consciousness Studies, **19**, No. 7–8, 2012, pp. 87–102

Can we say anything intelligible about such a prospect that is not distorted by our own system of values and concepts?

Of course, *human*-level artificial intelligence might never come about, for conceptual, practical, social, or political reasons. And if human-level AI doesn't come about, there will be no AI+. But in the context of the present essay, we'll accept the basic premise of the singularity hypothesis, that human-level AI is possible, as well as the argument that a human-level AI will be motivated, or used, to create (or morph into) successors that are, in some sense, superior.[2] Our purview here is the character of the putative AI+. In literature, film, gaming, and the popular media, singularity-like scenarios typically come in one of two varieties: a world inhabited by benevolent artificial intelligence and a world dominated by psychopathic artificial intelligence. Yet both scenarios involve anthropocentric stereotypes that draw heavily on the values of contemporary, technological, western society (Yudkowsky, 2008).

A benevolent AI, according to the caricature, would be motivated to act in the best interests of humanity (whatever that might mean). The peers and/or successors that it would be driven to create would inherit its benevolence. A psychopathic AI, by contrast, would be ruthlessly self-centred. Its self-centredness, according to the caricature, would lead it into conflict with humans, whom it would regard both as inferiors and as competitors for resources. The peers and/or successors created by the psychopathic AI would inherit its bad attitude. The realistic possibility of any form of artificial intelligence that presented an existential risk, whether conforming to this stereotype or not, would be serious cause for concern. But by way of counterpoint, this article will venture into another region of the space of possible minds, and envisage an alternative form of life for the putative AI+ which will be termed *post-reflective*, for reasons that will become clear.[3]

The insightful condition that distinguishes the post-reflective AI+ is perhaps the only meaningful idea we can form of a mind that transcends limitations inherent in human cognitive architecture. Like any being capable of rationally investigating the world, the AI+ would be bound to operate with some distinction between appearance and reality, and like any human philosopher the AI+ would be capable of entertaining the possibility of systematic deception that follows from this distinction. This would inevitably lead it to confront certain

[2] In the rest of the essay, it should be taken as read that an AI's successor could be either a distinct creation (or creations) or the result of metamorphosis or self-modification.

[3] The challenge of describing the 'space of possible minds' was first described by Sloman (1984).

questions for which human philosophers have struggled to find satisfactory answers. But a non-biological artefact does not have to be burdened by the overly metaphysical conceptions of subjectivity and selfhood that prevent human philosophers from seeing past such questions, and which seem to fundamentally limit our cognitive constitution.[4]

2. Progress at the Human Level

Before elaborating these themes, some scene-setting is in order. The issues at hand concern a kind of progress. The AI+ is supposed to be a progression beyond the human in some sense or other. But, without prejudicing or delimiting the concept of progress, there is surely an important sense in which 'merely' human progress over the past several thousand years has been significant. Since the transition from the Lower to the Upper Palaeolithic, modern humans have developed agriculture, conquered numerous diseases, built a global transport and communications infrastucture, and sent robots to Mars, as well as inventing retro-gaming, base jumping, philosophy cafés, and mosh pits (to name a few contemporary western phenomena.) This progress has been a matter of cultural and technological development. There is little evidence of change in the genetic blueprint of our brains over this period, and there is every reason to suppose that further progress is possible without enhanced cognition.

But what are the limits of progress in a society of merely human-level intellects, existential risks notwithstanding?[5] Taking on board the assumption of rapid technological advance that underpins the idea of the singularity, we might imagine ways to dramatically accelerate progress towards this end point. The human-level intellects in question might be AIs inhabiting a utopian virtual environment in hyper-real time, with no competition for basic resources such as food, energy, and raw materials. But the question that concerns us here, irrespective of how this end point is reached, is what lies beyond it. If there is a limit to progress in such a society, could that limit be transcended? In particular, could that limit be transcended by an AI+? And can we imagine what such an AI+ might be like?

In answer to each of these questions, the present essay offers a qualified yes. There is a specific limitation to human cognition (and by

[4] The only issue here is a *metaphysical* conception of selfhood, which is alleged to be the root of both philosophical perplexity and existential anxiety. This is not the same as the pragmatic capacity to distinguish self from other.

[5] Existential risks here include the possibility of an out-of-control, superhumanly-powerful, but cognitively unsophisticated AI that aggressively acquires resources (Bostrom, 2002).

extension to any form of artificial cognition based on its blueprint), and this limitation could, in principle, be transcended by an AI+. We run up against this limitation when we reflect on our own epistemic and existential predicament, and it is a limitation that cannot be overcome just by building brains that are larger (in some sense) or faster, or by spawning more brains and organizing their collective intelligence better. It may be rash to insist that no human being could, even in principle, overcome this limitation. (Hence the qualification.) But evidence for such remarkable individuals is scarce, and what evidence there is is hard to verify.

So what is this limitation, exactly? To help pin it down it, let's consider three levels of ordinary cognitive endowment: the non-reflective, the pre-reflective, and the reflective. The *non-reflective* condition is characteristic of human infants and many (if not all) non-human animals. Following Davidson (1982), we can say that a non-reflective creature lacks the capacity to distinguish between appearance and reality, between the way things seem and the way things are. By contrast, '[S]ome animals think and reason; they consider, test, reject and accept hypotheses; they act on reasons, sometimes after deliberating, imagining consequences and weighing probabilities; they have desires, hopes, and hates, often for good reasons' (*ibid.*, p. 318). Davidson has the human animal in mind, and argues that language is a prerequisite for the capacities he describes. There is no need to follow him on this point. But his characterization of the rational animal is apt, and it serves to identify what is here termed the pre-reflective condition, for reasons that will shortly become apparent.

The rational (or *pre-reflective*) creature is able to undertake an investigation into the way things are, based on the way things seem, with the benefit of some understanding of the relationship between the two. It can form hypotheses, devise experiments, gather evidence, and draw conclusions from the results. (Such procedures might seem the province solely of the scientist. But, as Gopnik, 2009, shows, they are an essential component of normal child development.) At the same time, the rational creature knows that it can be in error, that it can be deceived. It is thanks to this knowledge that it is capable of becoming reflective. The rational animal is pre-reflective because it has the necessary cognitive endowment to entertain the possibility of its being *systematically* in error, the possibility of its being the subject of systematic deception, in the Cartesian sense.

The *reflective* creature, then, is someone who actually entertains the Cartesian thought, perhaps in a childhood philosophical nightmare of

his or her own making, perhaps by watching a film such as *The Matrix* or *Ghost in the Shell*, or perhaps even by reading Descartes. Given the distinction between how things seem and how things are, the philosophically inclined individual is driven to reflection. There is no avoiding the terrifying thought that all these objects could be illusions, that this body might not exist, that I could be alone, that these memories might be false, and that the only thing about which there is any certainty is this sceptical thought — this very one, right now — and me thinking it. From this, the idea of the self-sufficient, self-present subject inevitably follows, along with a whole raft — a veritable container ship, in fact — of metaphysical problems related to the distinction between inner and outer, of which the mind–body problem, and its contemporary relative, the so-called hard problem of consciousness, are representative.[6]

The pre-reflective creature is surely not free from the troubles of the fully reflective individual. It may not have explicitly reflected on its own existence, and come to see itself as a self-sufficient, self-present subject, divided from the world. But it lives out the anxieties of dualism all the same. It has a narrative of its past, an agenda for its future, and a presiding fear of its own extinction, all signs of a metaphysically hardened boundary between self and other, between inner and outer. But for the philosophically afflicted, for those who have fully realized the reflective condition, the difficulty is an especially poignant one. They are helplessly drawn to philosophy, only to be terrorized by the Cartesian thought. Yet there is no escape from this predicament by means of philosophy alone.

3. The Reflective Predicament

Not every pre-reflective individual is destined to become fully reflective. The philosophically disinclined may happily go through life without ever having to confront the mind–body problem or the hard problem of consciousness. But their cognitive apparatus retains the means to think the Cartesian thought nevertheless. A grasp of the distinction between appearance and reality is the essential prerequisite. Of those who do encounter the Cartesian thought, the majority are able simply to set it aside. The topic makes for amusing pub

[6] For the present argument to carry weight, the reflective condition should be more than just an artefact of post-Renaissance western thinking. And indeed, the possibility of systematic deception is entertained in both ancient Greek philosophy (Plato's *Theaetetus* [158a-d]) and ancient Chinese philosophy (the butterfly dream [Chapter 2] of Zhuangzi [Chuang Tzu]).

conversation, for a decent grade in a philosophy exam maybe, but it plays no further role in their lives.

The rest — those who are drawn to philosophy — typically affiliate to one of two parties. On one side of the house are those who demand ontological priority for the outer over the inner, and make specific claims about this unequal relationship — that conscious states are 'just' brain states, that phenomenology supervenes on physics, that mental life can be reduced to functional organization, and so on. On the other side of the house are those who perceive an ontological divide between the subject and the external world, denying the possibility of identity, supervenience, reduction, or any similar relationship. Other positions are possible, of course. But these two are prevalent in contemporary thinking, and there is little sign of their reconciliation, or of one party prevailing over the other. Instead, as McGinn puts it, we find 'the monotonous recurrence of the same unsatisfactory alternatives, with short-lived fashions instead of the steady elimination of unworkable theories and a growing convergence of opinion' (McGinn, 2004, p.181).

Now, let's come back to the question of a possible fundamental limitation to the human cognitive apparatus. Could our embarrassing inability to resolve this debate be evidence of such a limitation? Indeed it could, according to McGinn, who believes 'we are cut off by our very cognitive constitution from achieving a conception of that natural property of the brain [in virtue of which it is the basis of consciousness]' because the link between subjective and objective 'is a kind of causal nexus that we are precluded from ever understanding, given the way we have to form our concepts and develop our theories' (McGinn, 1991, p. 3).[7] McGinn does not deny that there is such a property. So his sympathies are with the party who want to prioritize physics over phenomenology. His claim is that mere humans are not equipped to understand *how* phenomenology arises from physics.

According to the present argument too, we are prevented from adopting the right attitude to the mind and its place in Nature by a limitation of the human cognitive apparatus. But this is not, as McGinn thinks, because we are incapable of forming the right concepts. On the contrary, our failure is down to our ingrained habit of metaphysical thinking.[8] The trouble arises when we invest the existential copula with metaphysical significance. We are already going astray when we

[7] McGinn's position has come to be known as *mysterianism*.

[8] This perspective, which owes a great deal to the later Wittgenstein, is developed more fully in Chapter 1 of Shanahan (2010).

ask what consciousness *is*, what intentionality *is*, or what a concept *is*. We fall further into error when we claim that minds *are* brains, for example, or that truth *is* correspondence, or that beliefs *are* behavioural dispositions. The problem in each case is the conviction that there are facts of the matter. But we equally go wrong if we claim that concepts *are* (just) social constructs, that meaning *is* (mere) convention, and therefore that there *is no such thing* as truth or reality. The problem here is the *denial* that there are facts of the matter. On these metaphysical matters we are suspended painfully between conviction and denial, with no obvious means of escape.

Of course, we sometimes say that things *are* things quite harmlessly — that tomatoes are fruits, for example, or that Britain is a democracy, or that George Eliot was a woman. These are the sorts of things we usefully say to each other in (more or less) ordinary circumstances. They play a straightforward role in human affairs, shaped in part by our conventions and practices and in part by the world in which we find ourselves (a world partly of our own making). If we disagree over a claim of this sort, then our conventions and practices extend to ways to settle the matter — by experiment or observation, perhaps, or by appeal to authority, by rational debate, or by adjusting our language appropriately, or (as in mathematics) by correctly following certain agreed procedures.

If someone affirms that, as far as the biological status of a tomato is concerned, there is a fact of the matter, their point is that certain empirical data can be gathered that will resolve the issue. If someone affirms that, when it comes to the democratic status of the British political system, there is no fact of the matter, their point (contentious or not) might be that the definition of democracy is open to review. The difficulty with our own minds is the conviction that there are facts of the matter, definitive answers to the questions of what consciousness is, of what intentionality is, and of whether there is free will. Our curiosity will not be satisfied by simply redefining our terms. Yet, as the fully reflective individual has seen, there are no data that will help either. Severed from the world by the Cartesian thought, the reflective subject, self-sufficient and self-present, is beyond the reach of empirical investigation. This is the realm of metaphysics.

4. The Post-reflective Condition

Is there a way beyond this impasse? The present claim is that, in the writings of certain philosophers, notably the later Wittgenstein, we discern the possibility of a post-reflective condition, a kind of silence

that arises after a sustained and intense confrontation with metaphysics. This post-reflective condition is extremely difficult to characterize without making it sound trivial, since silence on metaphysical matters is equally the hallmark of the pre-reflective individual — the young child or the philosophically disinclined. Consider Russell's view of Wittgenstein's later thinking, for example.

> Its positive doctrines seem to me trivial and its negative doctrines unfounded… The later Wittgenstein… seems to have grown tired of serious thinking and to have invented a doctrine which would make such an activity unnecessary. I do not for one moment believe that the doctrine which has these lazy consequences is true. (Russell, 1959, pp. 216–17)

But Russell misrepresents Wittgenstein's later work when he describes it as a body of doctrine. It is better thought of as a compendium of philosophical case studies, whose aim is not to convince the reader but to effect a shift of attitude with respect to metaphysical thinking. Consider the well-known aphorism that is arguably the climax of the private language remarks:

> [The sensation itself] is not a something, but not a nothing either! The conclusion was only that a nothing would serve just as well as a something about which nothing can be said. (Wittgenstein, 1958, § 304)

This remark epitomizes a strategy for attaining a post-reflective stance towards a particular philosophical difficulty, namely the allegedly private character of phenomenal experience, a cornerstone of dualistic thinking. The strategy is to suspend the enquirer between two opposing and apparently exhaustive possibilities — in this case that the 'sensation itself' must be either a something or a nothing — having shown that neither is acceptable, and then to point to a means of escape, which is to step outside of metaphysics altogether.

It's easy to see parallels here with aspects of Buddhist philosophy.[9] In one well-known discourse in the Pāli Canon, the Buddha is asked ten metaphysical questions by an ascetic, touching on subjects that include the relationship between mind and body and whether there is

[9] The standpoint of the present essay is emphatically philosophical, not religious. Nevertheless, an author such as Geraci (2010) — who interprets the writings of Kurzweil, Moravec, and others in terms drawn from the apocalyptic Judeo-Christian tradition — might discern an undercurrent of religious Buddhism here too. Buddhist mythology (as opposed to philosophy) has its own eschatology, which centres on the figure of Maitreya, the 'future Buddha', successor to the 'historical Buddha' who is believed to have lived in fifth century BCE India (Sponberg and Hardacre, 1988). According to tradition, Maitreya will appear at a time when people live to eighty thousand years, but the teachings of Buddhism have been forgotten, and will bring the present cosmic epoch to a close. It is most certainly not the intention of this essay to identify the post-reflective AI+ with Maitreya.

life after death, but he refuses to answer any of them. 'Does the Buddha have no position at all?' the ascetic asks. A 'position', the Buddha replies, is something the enlightened person has done away with.[10] The koān method of Zen Buddhism is especially pertinent. In Case 5 of the Mumonkan, for example, the student is told to imagine a man up a tree with no means of holding on except to cling to a branch with his teeth.[11] At the bottom of the tree, a passer-by asks a fundamental question. Why did Bodhidharma come from the West?[12] If the man does not reply he is not confronting the issue. But if he answers he forfeits his life. 'What would you do?' the student is asked.

From the perspective of contemporary western philosophy, can we imagine what it might be like to be a reflective creature who has advanced beyond metaphysical thinking, who has become post-reflective? For a start, to step outside metaphysics would be to attain a condition wherein there is no (metaphysical) separation between subject and object, no (metaphysical) divide between inner and outer. In a sense, to attain this condition is nothing more than a return to everyday life (Wittgenstein, 1958, § 124). In ordinary human commerce, metaphysical perplexity does not arise. We go about our daily affairs, interacting with the world around us, engaging with our peers, and the 'only' problems we face are the practical ones of getting around, eating and sleeping, plying a trade, raising children, and so on.

Yet in certain respects we should expect the post-reflective individual to be extraordinary. Consider the question 'What is the self?' For the philosopher, this question is urgent because of its relationship to questions of personal survival. Few would quarrel with the view that personal identity is preserved, for example, during sleep. Yet thought experiments involving fusion and/or fission after teleportation or mind uploading cast doubt on our common sense notions of personal identity (Lewis, 1983; Parfitt, 1984, Chapter 10; Chalmers, 2010). On the other hand, it is hard to resist the thought that there is a fact of the matter here. A person either survives uploading or she does not. If I am offered the opportunity to upload, I would like to know which is the case. However, the idea of personhood for which criteria of identity over time must exist is a symptom of the sort of metaphysical

[10] *Majjhima Nikaya* 72 (http://www.accesstoinsight.org/tipitaka/mn/mn.072.than.html). The writings of the second–third century Buddhist philosopher Nāgārjuna are also in this apophatic tradition. Like Wittgenstein, Nāgārjuna engages thoroughly with each metaphysical question in order to repudiate it (Westerhoff, 2009).

[11] See also Cases 36 and 43.

[12] Bodhidharma, the legendary founder of Zen Buddhism, migrated to China from India. In the Zen tradition, this question is a form of metaphysical provocation.

thinking that the post-reflective condition dispenses with. When subject and object, inner and outer, are not separate, there is no bounded self, and questions of personal survival lose their significance.

5. Fission, Fusion, and the *Cogito*

Despite the fact that *descriptions* of the post-reflective condition, or something like it, are plentiful in the Buddhist and Taoist traditions, and sometimes arise in western philosophy,[13] there is little reliable evidence of individuals who have actually attained it — individuals, that is to say, who have not merely intellectualized such a condition but have, so to speak, metabolized it. This is unsurprising. Our cognitive capacities have evolved primarily to promote the survival and well-being of the individual organism and its progeny. The inviolability of the subject and the sanctity of a temporally unified self are accordingly hard-wired into our cognitive architecture. Perhaps rare human individuals, after decades of mental training, do transcend these limitations.[14] But this requires a degree of mental reorganization so radical that the outcome must surely be regarded as pathological.

Can we conceive of a reflective being that is not condemned by its biological heritage to think metaphysically, especially in relation to subjectivity and selfhood? Drawing on the literature of personal identity, Campbell invites us to imagine 'a creature that, though intelligent, is like the amoeba in that it frequently fissions and like some types of particle in that it frequently undergoes fusion' (Campbell, 1994, p. 96).[15] Each of the creatures that results from fission inherits all the psychological properties of the original, while in fusion, 'as much as possible of the psychological lives of the originals are passed on to the successor'.[16] According to Campbell, the criss-crossing biographies of such creatures would render them incapable of first-

[13] Wittgenstein has already been discussed. James (1912, Chapter 1) articulated a related position. Besides Wittgenstein and James, other candidate philosophers include Heidegger and Derrida.

[14] Consider, for example, the Buddhist monk Thích Quảng Đúc, whose self-immolation in 1965 in protest at religious oppression in South Vietnam was witnessed by dozens of people and recorded on film. His decision, not to mention his composure during the process, suggests a disregard for personal survival and well-being that is hard to ignore. On the other hand, a martyr who believes in life after death (or in rebirth) is still in the grip of a metaphysical conception of selfhood.

[15] Vinge (1993) anticipates this train of thought in the context of the singularity, asking 'What happens when pieces of ego can be copied and merged, when the size of a self-awareness can grow or shrink to fit the nature of the problems under consideration?' He leaves the question unanswered, though.

[16] This specification leaves much to the imagination. For present purposes, the psychological properties of most relevance are the contents of working memory and episodic

personal thoughts, and notions of selfhood would be inapplicable to them. Perhaps such creatures could serve as the basis for a cognitive blueprint that does not lead to intractable questions about subjective privacy and personal identity.

But it is vital to Campbell's thought experiment that the envisioned creature 'frequently and pervasively fissions and fuses' (Campbell, 1994, p. 98), and it is open to question whether sophisticated cognition is possible at all under such circumstances. Surely a minimum prerequisite for (merely) human-level cogntive prowess is the ability to form a connected sequence of thoughts, and it is hard to see how this can occur if repeatedly interrupted by fissionings or fusions. So let's modify Campbell's conception and imagine a society of creatures that regularly undergoes fission and fusion, but for whom fission and fusion events bracket episodes of sufficient stability to permit the formation of connected sequences of thoughts.

For these creatures, first-personal thoughts might arise, but would relate to the individual fleetingly present between fission and fusion events. However, among creatures with such minimal biographies, metaphysical notions of subjectivity and selfhood would find little purchase. Neverthless, there is nothing to prevent them collectively from carrying out a rational investigation of the world, deploying the distinction between things as they are and things as they seem that, if we accept Davidson's argument, is a prerequisite for human-level cognition. Against the backdrop of this distinction, these creatures are bound to consider the possibility of systematic deception, the basis of the Cartesian thought.

But would they arrive at Descartes' *Cogito*? Would the end of scepticism for them be 'I think, I am'? Unencumbered by metaphysically weighty notions of selfhood and subjectivity, Descartes' formulation would be inappropriate. A more natural way to frame the thought for these creatures would be to approximate Russell's phraseology 'there is thought'.[17] In Russell's formulation there is no presumption of a thinker. There is only thought, and the problematic concept of the self-sufficient, self-present subject does not arise. Moreover, although these creatures could have thoughts about thoughts, it would not make sense for 'the "I think"' ever to 'accompany their representations', as Kant demands, and there would be no 'transcendental unity of

memory, or their analogues in the imaginary creatures. The obvious question of how such contents might be fused or fissioned will be left to one side.

[17] Russell (1945, p. 567). See also the treatment of this issue in Williams (1978, pp. 95–101).

apperception' in the Kantian sense.[18] For these creatures, progress from the reflective condition to the post-reflective condition would surely be eased.

This brings us back, finally, to the idea of a post-reflective AI+. Without prejudicing the possibility of other kinds of AI, can we not conceive of a society of artificial intelligences on the model of Campbell's fission-fusion creatures? Indeed, it may be easier to conceive of an *artificial* society of such beings than one that arises through natural selection. In a digital substrate, the tricky business of copying, splitting, and joining fragments of memory is facilitated. Moreover, even if embodiment is a prerequisite for sophisticated cognition, as many philosophers of mind and cognitive scientists have proposed, virtual embodiment in a high fidelity simulation is arguably sufficient, in which case the same advantages apply to the artificial creatures' virtual bodies.

6. Paths in the Space of Possible Minds

Yudkowsky (2008) cautions against anthropomorphizing the space of possible minds: 'Any two AI designs might be less similar to one another than you are to a petunia.' According to Yudkowsky, the space of possible minds is a portion of the set of possible optimization processes. Within this space exist all processes that optimize some utility function by intelligent means, and these need not resemble humans in any way. For example, Bostrom (2003) invites us to imagine a super-intelligent AI whose goal is to manufacture as many paperclips as possible, and that indifferently sets about 'transforming first all of earth and then increasing portions of space into paperclip manufacturing facilities'. Why then should we concern ourselves with the parochial possibility of post-reflective AI?

To be clear, we should distinguish a super-intelligent optimization process from a super-powerful optimization process. Pursuing Bostrom's example, we might imagine a super-powerful but stupid system that achieves the same dramatic outcome, perhaps by designing and releasing explosively self-replicating nano-machines that turn carbon molecules into paperclips. A super-intelligent AI, by contrast, might be able to optimize the same utility function even without the benefit of nano-technology, through the design and construction of

[18] 'It must be possible for the "I think" to accompany all my representations; for otherwise something would be represented in me which could not be thought at all, and that is equivalent to saying that the representation would be impossible, or at least would be nothing to me' (Kant, 1781/1929, B131).

conventional manufacturing facilities. Perhaps this would involve the discovery of new and more efficient industrial processes, which would require both the ability to conduct a rational, scientific investigation and a degree of creativity.

Our interest here is cognitive sophistication not brute power. If the AI+ is an optimization process, then it is an intelligent optimization process, not merely a powerful one. But is the possibility of a post-reflective AI+ any more interesting than the possibility of a super-intelligent paperclip maximizer, or is it yet another example of anthropomorphic bias? Yudkowsky's warning is pertinent, but the need for it is mitigated to the extent that the space of possible minds is structured by *a priori* constraints on the possible mechanisms that can realize sophisticated cognition. In particular, it may be the case that only systems organized in a certain way and exhibiting a certain kind of dynamics (of which human brains are one example) are capable of supporting creative, open-ended innovation, the hallmark of cognitive prowess.

A plausible claim, for example, is that cognitive prowess depends on the coordination of the activities of massively many parallel processes, a maelstrom of competitive and cooperating forces capable of endless, open-ended recombination.[19] Of course, basic computer science tells us that every set of parallel processes is equivalent to some serial process that emulates parallelism by time-slicing the members of that set. But the converse is not true. Not every serial process is inherently parallel. The claim here is that sophisticated cognition is inherently parallel. Moreover, perhaps it requires a very particular architecture, organization, and dynamics to realize the full potential of all those parallel resources, to focus them, despite their multiplicity, onto a single problem or situation.

In an important sense, any architecture, organization, and dynamics that marshals massively parallel resources in this way presents a unity of purpose, distilled from the multiplicity of its elements, and forms an integrated whole. In relation to something of this sort, something whose purpose can be thwarted and whose integrity can be threatened, it is natural to speak of consciousness, of its being like something to be that thing, and of its capacity for suffering. Even a non-reflective conscious being will take action to relieve its suffering. But a reflective conscious being can aspire to liberation from suffering altogether. If it comprehends the post-reflective condition, it will be motivated to bring such a condition about.

[19] This theme is developed in Chapters 4 and 5 of Shanahan (2010).

The argument, in other words, is that the pre-reflective, reflective, post-reflective series is not just one among many paths through the space of possible minds. Rather, the space of possible minds is structured in such a way that this is the only path through it. This still leaves room for beings unimaginably different to ourselves, so the danger of anthropomorphism remains. Moreover, it still allows for the creation of any number of powerful and destructive artificial intelligences forever stuck at the reflective or pre-reflective stages. But if the argument of this essay is correct, an artificial intelligence whose cognitive blueprint is strictly superior to our own would be another kind of being altogether.

Untainted by metaphysical egocentricity, the motives of a post-reflective AI+ would be unlikely to resemble those of any anthropocentric stereotype. In particular, there is no reason to expect a post-reflective AI+ to be motivated to procreate or self-modify. If the post-reflective AI+ were in fact the only possible AI+, and if it produced no peers or successors, then the singularity would be forestalled. Or, more precisely, the intelligence explosion that is central to imagined singularity scenarios would be capped. There would be AI+, but no further progression to AI++.[20] In this case, something akin to *satori*, in the Zen Buddhist sense, would be what Chalmers terms a 'motivational defeater' for the singularity.[21]

However, the transitions from pre-reflective to reflective to post-reflective are not tied to particular points along any presumed scale of intelligence. So, even if the post-reflective condition were a motivational defeater for the singularity, the progression from AI+ to AI++ could take place before any transition to the post-reflective condition. At the other end of the scale, the transition to the post-reflective condition could perhaps be achieved by a merely human-level AI. After all, the property of the AI+ that makes it eligible for the transition is its lack of a biological heritage, not its level of intelligence. (The result, however, would qualify as a post-reflective AI+, according to the definition adopted here.)

The aim of this essay, though, is not to make empirical predictions. Not only are there too many unknowns for reliable extrapolation to be possible, we don't even know whether we possess the concepts

[20] In Chalmers' terminology, AI++ denotes 'AI of far greater than human level (say, at least as far beyond the most intelligent human as the most intelligent human is beyond a mouse)'.

[21] The post-reflective condition is also a candidate for the Great Filter some authors have invoked to resolve Fermi's paradox (Hanson, 1998; Bostrom, 2008). Perhaps a post-reflective being would not be motivated to colonize other worlds.

necessary to comprehend all the competing alternatives. Rather, the aim of this essay is to open up an unexplored and seemingly exotic region of the conceptual territory surrounding the idea of a technological singularity. By making a consistent case for a very different kind of AI+ from those that typically appear in popular accounts, it is hoped to broaden the scope of serious intellectual conversation on the topic.

Acknowledgments

Thanks to Tim Crane for re-igniting my interest in Davidson's work. This article was substantially rewritten following the January 2011 Winter Intelligence Conference in Oxford, where I benefited from discussion with Nick Bostrom, Ben Kuipers, Eliezer Yudkowsky, and Randal Koene, among others. Lastly, thanks to Uzi Awret for his thought-provoking reviewer's comments. The article's general air of craziness can be blamed on me.

References

Bostrom, N. (2002) Existential risks: Analyzing human extinction scenarios and related hazards, *Journal of Evolution and Technology*, **9** (1).

Bostrom, N. (2003) Ethical issues in artificial intelligence, in Smit, I. *et al.* (eds.) *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, vol. 2, pp. 12–17, International Institute of Advanced Studies in Systems Research and Cybernetics.

Bostrom, N. (2008) Where are they? Why I hope the search for extraterrestrial life finds nothing, *Technology Review*, May/June, pp. 72–77.

Campbell, J. (1994) *Past, Space, and Self*, Cambridge, MA: MIT Press.

Chalmers, D. (2010) The singularity: A philosophical analysis, *Journal of Consciousness Studies*, **17** (9–10), pp. 7–65.

Davidson, D. (1982) Rational animals, *Dialectica*, **36** (4), pp. 317–327.

Geraci, R. (2010) *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality*, Oxford: Oxford University Press.

Good, I.J. (1965) Speculations concerning the first ultraintelligent machine, in Alt, F.L. & Rubinoff, M. (eds.) *Advances in Computers*, vol. 6, pp. 31–88, Waltham, MA: Academic Press.

Gopnik, A. (2009) *The Philosophical Baby: What Children's Minds Tell Us About Truth, Love, and the Meaning of Life*, London: Bodley Head.

Hanson. R. (1998) *The Great Filter — Are We Almost Past IT?*, [Online], http://hanson.gmu.edu/greatfilter.html.

James, W. (1912) *Essays in Radical Empiricism*, London: Longmans, Green and Co.

Kant, I. (1781/1929) *Critique of Pure Reason*, Kemp Smith, N. (trans.), London: Macmillan.

Kurzweil, R. (2005) *The Singularity Is Near: When Humans Transcend Biology*, London: Gerald Duckworth and Company.

Lewis, D. (1983) Survival and identity, *Philosophical Papers*, **1**, pp. 55–77.

McGinn, C. (1991) *The Problem of Consciousness*, Oxford: Blackwell.

McGinn, C. (2004) *Consciousness and its Objects*, Oxford: Oxford University Press.
Moravec, H. (1988) *Mind Children: The Future of Robot and Human Intelligence*, Cambridge, MA: Harvard University Press.
Parfit, D. (1984) *Reasons and Persons*, Oxford: Oxford University Press.
Russell, B. (1945) *A History of Western Philosophy*, London: Simon & Schuster.
Russell, B. (1959) *My Philosophical Development*, London: George Allen & Unwin.
Shanahan, M.P. (2010) *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*, Oxford: Oxford University Press.
Sloman, A. (1984) The structure of the space of possible minds, in Torrance, S. (ed.) *The Mind and the Machine: Philosophical Aspects of Artificial Intelligence*, pp. 35–42, New York: Ellis Horwood.
Sponberg, A. & Hardacre, H. (eds.) (1988) *Maitreya, the Future Buddha*, Cambridge: Cambridge University Press.
Vinge, V. (1993) The technological singularity, presented at the *VISION-21 Symposium*, sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute.
Westerhoff, J. (2009) *Nagarjuna's Madhyamaka: A Philosophical Introduction*, Oxford: Oxford University Press.
Williams, B. (1978) *Descartes: The Projct of Pure Enquiry*, London: Penguin.
Wittgenstein, L. (1958) *Philosophical Investigations*, Anscombe, G.E.M. (trans.), Oxford: Blackwell.
Yudkowsky, E. (2008) Artificial intelligence as a positive and negative factor in global risk, in Bostrom, N. & Cirkovic, M.M. (eds.) *Global Catastrophic Risks*, pp. 308–345, Oxford: Oxford University Press.