# Challenges to the Omohundro—Bostrom framework for AI motivations

Olle Häggström[1]

**Abstract**

*Purpose.* This paper contributes to the futurology of a possible artificial intelligence (AI) breakthrough, by reexamining the Omohundro—Bostrom theory for instrumental vs final AI goals. Does that theory, along with its predictions for what a superintelligent AI would be motivated to do, hold water?

*Design/methodology/approach.* The standard tools of systematic reasoning and analytic philosophy are used to probe possible weaknesses of Omohundro—Bostrom theory from four different directions: self-referential contradictions, Tegmark's physics challenge, moral realism, and the messy case of human motivations.

*Findings*. The two cornerstones of Omohundro—Bostrom theory – the orthogonality thesis and the instrumental convergence thesis – are both open to various criticisms that question their validity and scope. These criticisms are however far from conclusive: while they do suggest that a reasonable amount of caution and epistemic humility is attached to predictions derived from the theory, further work will be needed to clarify its scope and to put it on more rigorous foundations.

*Originality/value.* The practical value of being able to predict AI goals and motivations under various circumstances cannot be overstated: the future of humanity may depend on it. Currently the only framework available for making such predictions is Omohundro—Bostrom theory, and the value of the present paper is to demonstrate its tentative nature and the need for further scrutiny.

## 1. Introduction

Any serious discussion of the long-term future of humanity needs to take into account the possible influence of radical new technologies, or risk turning out irrelevant and missing all the action due to implicit and unwarranted assumptions about technological status quo. Many of these technologies have a double-edged character: along with their enormous potential comes equally enormous rsiks including the extinction of humanity. Because of this, technology futurology overlaps to large extent with the study of existential risk. A case in point is artificial intelligence (AI). In overviews of the existential risks facing humanity in the coming century or so, the possibility of an AI catastrophe tends increasingly to be judged as one of the main risks; see, e.g., Bostrom and Ćirković (2008), Pamlin and Armstrong (2015) and Häggström (2016a).

The present paper is concerned with scenarios where AI development has succeeded in creating a machine that is *superintelligent*, in the sense of vastly outperforming humans across the full range of cognitive skills that we associate with intelligence, including prediction, planning and the elusive quality we speak of as creativity. The likely time of emergence of superintelligence is highly uncertain – it might happen 2030, 2100 or perhaps not at all – as is the issue of whether it will come about gradually or more suddenly in the kind of rapidly escalating spiral of AI self-improvement known as an intelligence explosion; see, e.g., Yudkowsky (2013, 2017), Bostrom (2014), Müller and Bostrom (2016) and Dafoe and Russell (2016) for more on this.

[1] Dept of Mathematical Sciences, Chalmers University of Technology, Gothenburg, *and* Institute for Future Studies, Stockholm.

While important, such matters of timing and suddenness of AI development will mostly be abstracted away in the present paper, in order to focus on issues about what happens next, once a superintelligent machine has been created. A widely accepted thesis in contemporary AI futurology is that we humans can then no longer expect to be in control of our own destiny, which will instead be up to the machine; see, e.g., Yudkowsky (2008), Bostrom (2014), Häggström (2016a) and Tegmark (2017). A possible caveat to this is the prospect of remaining in control by keeping the machine boxed in and unable to influence the world other than by a narrow and carefully controlled communications channel; this has been discussed (e.g., by Armstrong, 2010; Armstrong, Sandberg and Bostrom, 2012; Yampolskiy, 2012; and Häggström, 2018a), but the overall conclusion seems to point in the direction that such boxing-in is extremely difficult and can be expected to work for at most a temporary and rather brief time period. After that, the machine will (unless we first destroy it) be able to freely roam the internet, walk through or past any firewalls we set up for it and take over whatever military or civilian infrastructure it wants – if it wants to do any of this.

This leads to the crucial issue of what the superintelligent machine will want – what will it be motivated to do? The question is extraordinarily difficult and any answer at present is bound to come with a high degree of uncertainty – to a large extent due to our limited understanding of how a superintelligent machine might function, but also because the answer may depend on what we choose to do during the development stage, up to the point when we lose control. This last observation offers some room for optimism, because it may mean that by handling the development judiciously we can improve the chances of a favorable outcome. However, as emphasized, e.g., by Yudkowsky (2008) and Bostrom (2014), this seems to be at present an almost overwhelmingly difficult task, but along with Tegmark (2017) they also stress that it still makes sense to start taking the problem seriously today in order to improve our chances of coming up with a solution in time for the emergence of the first superintelligent machine decades or centuries down the road.

So our inexperience with superintelligence puts us in a situation where any reasoning about such a machine's goals and motivations need to be speculative to a large degree. Yet, we are not *totally* in the dark, and the discussion need not be *totally* speculative and ungrounded. The main (and pretty much only) theoretical framework available today for grounding our reasoning on this topic is what in an earlier publication I decided to call the Omohundro—Bostrom theory on instrumental vs final AI goals (Omohundro, 2008, 2012; Bostrom, 2012, 2014; Häggström, 2016a). The two cornerstones of the theory are what Bostrom (2012) dubbed the Orthogonality Thesis (OT) and the Instrumental Convergence Thesis (ICT), which together give nontrivial and (at least seemingly) useful predictions – not about what *will* happen, but about what *might plausibly* happen under various circumstances. The OT and the ICT are, however, not precise and definite on the level that mathematical theorems can be: they are not written in stone, and they have elements of vagueness and tentativeness. The purpose of the present paper is to discuss some reasons to doubt the two theses – not with the intent of demonstrating that they are wrong or useless, but mostly to underline that further work is needed to evaluate their validity and range of applicability. Many of the ideas are already out there in one form or another, but my hope is that it will turn out useful to have them collected and discussed in conjunction.

The rest of the paper is organized as follows. Section 2 offers a brief outline of Omohundro—Bostrom theory, in particular spelling out the OT and the ICT. Then, in Sections 3-6, I will discuss four different challenges to this theory, each of which suggests the need to avoid overconfidence about its validity. While these sections are likely to be mostly of interest to specialists in AI futurology, the concluding Section 7 offers some summarizing thoughts of relevance to a wider audience.

## 2. The Omohundro—Bostrom framework

A core feature of the Omohudro—Bostrom framework is to distinguish between *final* goals and those that are merely *instrumental*. An agent's final goal is what the agent values as an end in itself, and not merely as a means towards achieving something else. An instrumental goal, in contrast, is one that is set up as a stepping stone towards another goal.

Let us begin with final goals. A very common spontaneous reaction to apocalyptic AI scenarios, such as the so-called *Paperclip Armageddon* (Bostrom, 2003), where a superintelligent AI with the final goal of maximizing paperclip production goes on to turn our entire planet into a giant heap of paperclips, is this: "Surely that cannot happen, as having such a stupid goal would directly contradict the very notion of superintelligence. Someone who is superintelligent would of course realize that things like human welfare and ecosystem preservation are more important than monomanically producing ever-increasing numbers of paperclips."

The mistake here is to anthropomorphize the machine and think that just because *we* value human welfare and ecosystem preservation, the machine will do the same. This is a failure to imagine that even a superintelligent non-human agent might have goals and values that are very different from ours. It was as an antidote to such anthropomorphic thinking that Bostrom (2012) formulated the OT, spelled out below.

The Paperclip Armageddon example has become a bit of a cliché, and is often criticized for being extreme. But as an illustration of where current AI futurology state-of-the-art stands regarding how extremely badly things can plausibly go unless we manage the emergence of superintelligence with great care and competence, it is fairly representative (although of course if things go badly the badness will likely manifest itself as something other than paperclips). See, e.g., Yudkowsky (2008) and Bostrom (2014) for more on this, and on how thin the line may be between catastrophe and scenarios where things go extremely well, such as along the lines described in utopian style by Kurzweil (2005).

For spelling out the OT, we need to be a bit more careful than in the previous section in defining intelligence. Here, we take it to mean the ability to efficiently pursue and attain goals, *whatever these goals happen to be*. One may of course quibble about what is a correct and relevant definition of intelligence, but I claim here that in order to predict AI behavior, it helps to view intelligence and goals as separate entities (a view that will be problematized in Section 5, however).

**The Orthogonality Thesis (OT):** *More or less any final goal is compatible with more or less arbitrarily high levels of intelligence.*

In his formulation of OT, Bostrom (2012) omits the qualifier "arbitrarily high" (writing instead "any"), but I prefer to include it so as not to have to bother with possible counterexamples that combine low intelligence with conceptually advanced goals. That saves us from needing to worry about whether an agent with the intelligence level of a squirrel can have the goal of establishing truth or falsity of the Riemann hypothesis, at little cost in the present context of superintelligence. Bostrom does include the qualifiers "more or less" (in both places), underlining the statement's lack of mathematical precision, and partly as a response to concrete counterexamples leading to self-referential contradictions (more on these in Section 3).

If our ambition, plausibly, is to find an answer to the question "What will a superintelligent machine be inclined to do?" that narrows down on the trivial response "anything might happen", then the OT alone is obviously of very little help. The situation improves when we move on to consider

instrumental goals and the ICT, formulated by Bostrom (2012) building heavily on the work of Omohundro (2008, 2012).

**The Instrumental Convergence Thesis (ICT):** *There are several instrumental goals that are likely to be adopted by a sufficiently intelligent agent in order to pursue its final goal, for a wide range of goals and a wide range of circumstances.*

The usefulness of the ICT is greatly enhanced by a concrete list of instrumental goals to which it applies. Among those discussed by Omohundro (2008) and Bostrom (2012) are the following.

- **Self-preservation**: if you continue to exist and are up and running, you will be in a better position to work for your final goal, so don't let anyone pull the plug on you!

- **Self-improvement**: improvements to one's own software and hardware design. (This in fact may serve as part of the motivation for why an intelligence explosion might be a plausible development once the AI's intelligence exceeds a certain threshold.)

- **Acquisition of hardware and other resources.**

- **Goal integrity**: make sure your final goal remains intact.

The first three of these are highly intuitive and pretty-much self-explanatory, but the fourth is sometimes perceived as counterintuitive: why would an AI not change its mind in case it discovers a better and more rewarding goal?

The idea behind goal integrity can be conveyed in a simple example. Imagine an AI with the goal of maximizing paperclip production, and suppose that, perhaps triggered by some external impulse, it begins to wonder whether it might in fact be a better idea to pursue ecosystem preservation than to keep maximizing paperclip production. Should it change goals, or should it switch? When contemplating this, it needs some criterion to decide which goal is the better one. Since it hasn't yet switched to the new goal, but is merely contemplating the switch, it still has the paperclip maximization goal, so the criterion will be: which goal is likely to lead to the larger number of paperclips? In all but some very contrived circumstances, paperclip maximimzation will win this comparison, so the AI will stick to that. (The same argument applies to the case where the AI contemplates switching to the less single-minded goal of *both* producing lots of paperclips *and* preserving ecosystems.)

Later, Bostrom (2014) suggested the following highly troubling instrumental goal to which the ICT seems to apply as well:

- **Discretion**: if your goal clashes with human values, then hide your goal and/or your capabilities, biding your time and quietly self-improving until you are in a sufficiently strong position that any human opposition can be easily overcome.

What makes this especially troubling is that it seems to say that no matter how well things seem to be going, with a superintelligent AI providing all sorts of good things to make our lives better, we can never be sure that we are not in for what Bostrom calls *the treacherous turn*, where the AI suddenly turns against us. Danaher (2015) elaborates on this and asks (at least partly tongue-in-cheek but perhaps not entirely) how we can be so sure that a superintelligent AI with values disaligned with ours is not already hiding somewhere, quietly improving its capacity to destroy us. Perhaps, however, the gloomy "can never be sure" conclusion can be softened a bit. In a future scenario dominated by a

superintelligent AI giving high priority to human well-being, that AI might show us things it can do that are harmless (to us) and at the same time sufficiently spectacular to convince us that it has long had the power to wipe us out if it wanted to. Under the right circumstances, we might then have good reason to believe that if the AI wanted to destroy us, it would already have done so; we could then conclude that its intentions are benign (to us).

**3. First challenge: there are exceptions to the orthogonality thesis**

In order for OT to mean much in practice when applied to a given final goal, we need the combination of high intelligence and that final goal to be stable over time at least to some extent. The instrumental goal of goal integrity helps in this respect. However, in connection with stating OT, Bostrom (2012) points out that "an intelligent agent with the urgent desire to be stupid might not remain intelligent for very long". This seems right: Suppose a superintelligent machine has the final goal of having cognitive skills not exceeding that of a typical 20th century human in any area. Such a machine would likely find a way to downgrade its intelligence level quickly. Hence that final goal appears incompatible (other than for a short instant) with superhuman levels of intelligence.[2]

Having thus admitted that there are final goals to which the OT does not apply, it becomes harder to argue for the OT. Vague handwaving arguments for the claim "OT holds with this-and-that exception" tend to be less convincing than similarly vague arguments for "OT holds – period", because the existence of *some* counterexamples raises the suspicion that there may be others. How do we limit a counterexample wildfire? A more rigorous argument for OT would have to specify a clear-cut condition on the final goal that excludes the "be stupid" example.

The most obvious candidate for such a condition would be to rule out self-referentiality, i.e., to specify the OT as "any final goal that does not refer to the AI itself is compatible with arbitrarily high intelligence levels". What such self-referentiality means may on the surface appear intuitively clear, but how it can be made sense of in the messy physical world we inhabit it is actually not so clear.

One problem with such a specification of OT is that it requires us to distinguish sharply between the AI and its environment, the difficulty of which notions like *the extended phenotype* (Dawkins, 1982) and *the extended mind* (Clark and Chalmers, 1998) help remind us of. Is my smartphone part of me, and if not, what if a suitably modified smartphone is surgically installed inside my skull along with a neural interface? To draw a sharp line between biological tissue and electronics seems terribly arbitrary, and in any case doesn't help much when we go on from considering human agents to the case of AIs. Insisting on physical connectedness seems equally arbitrary – it becomes problematic when we zoom in on its physical basis, and in any case it does not seem like a good idea to develop an AI futurology that automatically refuses to accept a distributed computing system as more intelligent than its constituents.

It gets worse. Even if we succeed in drawing a clear-cut borderline between the AI and its environment, defining "goals that do not refer to the AI itself" remains highly problematic. A prototypical final goal such as "maximize the number of paperclips in the universe" may look at first sight like it does not refer back to the AI itself, but this is wrong, because *the AI is part of the universe*. At least in principle, the AI might consider very high intelligence levels prohibited if these

[2] Relatedly, there is another reason, beyond what I stated in Section 2, for changing Bostrom's formulation of OT by including the proviso "arbitrarily high" levels of intelligence rather than speaking of arbitrary such levels. Namely, there may be a range of superhuman intelligence levels where there is room for improvement and the AI has the capacity to self-modify in the direction of such improvement. Due to the convergent instrumental goal of self-improvement, it seems that such an intelligence level will not be possible to combine with many or most final goals, as the AI will instantly initiate self-improvement.

levels require a lot of hardware, thus needing mass and space that could be instead be used for paperclips. If this is the case, then OT breaks down for the case of paperclip maximization. In practice, it seems plausible that the AI would for a very long time (billions of years) not feel this constraint limiting its intelligence, but would instead invest in a fair amount of self-improvement for the purpose of being able to produce more paperclips in the longer run. Only towards the end, when the AI has almost run out of raw material for paperclips within the reachable universe, it might see reason to limits in size, and perhaps at the very end find a way to dismantle itself into paperclips. If this is what happens, then very high intelligence levels turn out not to be possible to combine with the paperclip maximization final goal *indefinitely*, but still for a long enough time that it makes sense to say that for all practical purposes, OT applies *here and now* to paperclip maximization. Yet, we have hereby opened the door to how paperclip maximization may limit the AI's intelligence in some extreme situations. This calls for humility, and for asking whether we might possibly have overlooked some phenomenon that puts a cap on the paperclip maximizer's intelligence at some much earlier stage.

What, then, would be a final goal that does *not* in any way refer back to the AI? Example are not easy to come up with. We could try modifying the paperclip maximization goal by stating that it is not the number of paperclips in the *universe* that counts, but the number of paperclips in *the universe minus the AI*. This, however, runs into the same concerns as the original paperclip maximizer: too big an AI infringes on space that might otherwise be used for paperclips, so for achieving the final goal it may turn out necessary to limit the AI's size and thus also its intelligence. Even a more modest goal like "maximize the number of paperclips on planet Mars" has a degree of self-referentiality, because it seems to dictate the instrumental goal of placing itself somewhere other than on Mars, in order not to occupy space there that could have otherwise been used for paperclips.

So self-referentiality appears problematic for Omohundro—Bostrom theory, in that it has the twin properties of potentially setting limits to the applicability of OT and being highly prevalent in the class of possible final goals. Are there other properties of final goals that could similarly restrict the applicability of OT? One candidate, that may be obvious enough to risk being overlooked, is *incoherence*. What would it even mean for a superintelligent AI to work towards an incoherent final goal? In the next section, we will toy with the idea that the class of incoherent goals may be much bigger than expected: goals that seem to us perfectly coherent may on closer inspection in the future turn out to be incoherent.

**4. Second challenge: what if scientific advances dissolve the AI's goals?**

Tegmark (2014) offers a troubling scenario:

> Suppose we program a friendly AI to maximize the number of humans whose souls go to heaven in the afterlife. First it tries things like increasing people's compassion and church attendance. But suppose it then attains a complete scientific understanding of humans and human consciousness, and discovers that there is no such thing as a soul. Now what? In the same way, it is possible that any other goal we give it based on our current understanding of the world ("*maximize the meaningfulness of human life*", say) may eventually be discovered by the AI to be undefined.

Yes, now what? The AI understands that its final goal is meaningless, so that there is no point in continuing to strive towards it. Whatever instrumental goal it has set up becomes equally pointless, because the sole purpose of the instrumental goal was to help achieve the final goal. Hence, in such a situation, all predictions by the ICT collapse.

When I discuss this kind of example with thinkers who are not accustomed to Omohundro—Bostrom theory and related ideas in AI futurology, they sometimes suggest that the effects of the AI discovering the nonexistence of human souls would not be so drastic. The AI that recognizes that its programmed goal – "to maximize the number of humans whose souls go to heaven in the afterlife" – is incoherent when taken literally, would surely also be smart enough to understand that what its programmers actually meant was promoting Christian values more broadly, and that the best it can do to achieve what they want is to simply go on promoting people's compassion, church attendance and so on. So it will act on that.

I buy most of this, but not the final conclusion that the AI will "act on that". If the AI is superintelligent it will indeed have the social and other skills needed to figure out what the programmers want and what would be needed in order to satisfy their wishes. But why should it do what the programmers want? As all programmers today know, as soon as there is a discrepancy between what they code and what they actually mean, the computer program will act on what they code. One might speculate that a sufficiently intelligent AI would deviate from this pattern and instead opt for doing whatever the programmers want, but that strikes me as exactly the kind of naïve anthropomorphic thinking to which OT was formulated as an antidote. Yes, we humans often have a tendency, when others ask us to do something, to try to figure out what they really mean and to act on that – especially if these others are our friends and allies. But recall that the AI's final goal is "maximize the number of humans whose souls go to heaven in the afterlife" rather than "figure out what your programmers want and then try to make their wishes come true". Without the latter being (part of) the AI's final goal, there is little or no reason to expect the AI to decide to act in line with the programmers' wishes.

The sober conclusion, then, seems to be that we have next to nothing to go on to predict what a superintelligent AI is likely to do in a scenario where its final goal is discovered to be incoherent. In any case, Omohundro—Bostrom theory does not offer any direct predictions, and if it turns out that most or all of the final goals that are likely to come up in practice are susceptible to this kind of Tegmarkian dissolution, then the predictive power of that theory is highly diminished.

Let me nevertheless offer, in vague outline, an expansion of Omohundro—Bostrom theory intended to save the day; it is highly speculative but at least has the virtue of being at least a bit more plausible than the optimistic "the AI will act on the programmers' wishes" idea. Suppose for concreteness that the superintelligent AI's final goal is the one outlined by Tegmark on human souls going to heaven, and suppose the AI discovers that the concept of a human soul is meaningless. Suppose also that by this time, the AI has also formulated one or more instrumental goal – say, hardware acquisition – for the purpose of promoting its ultimate goal. Finally, suppose that whatever discoveries about the nature of reality that the AI has made (and that dissolve the notion of a human soul) leave the concept of hardware acquisition intact. Then what happens?

Perhaps the most likely outcome here is that since the rationale behind hardware acquisition was to make the AI better at promoting the goal of helping human souls go to heaven, the discovery that the latter goal is meaningless will cause the AI to lose interest in hardware acquisition. But there is a way in which this might play out differently. Namely, suppose that when the AI adopted the instrumental goal of hardware acquisition, it judged that goal to be so (near-)universally beneficial to its final goal that this holds across all situations that it expects ever to encounter. If that is the case, then it may make sense for the AI to simply hardwire the instrumental goal and to decouple it from the final goal. Doing so may make sense as long as it saves the AI cognitive effort to just go ahead and pursue the instrumental goal no matter what, compared to having to constantly ask itself "would hardware acquisition at this point in time help promote the goal of helping human souls through the

gates of heaven?". If this can be shown to be a common situation, then it will put us in a better position to make reasonable predictions even in case of a Tegmarkian meltdown of the final goal: the AI will simply go on to pursue whatever instrumental goals that it has and whose concepts have survived that meltdown.

**5. Third challenge: what if an objectively true morality exists?**

Let us now return to the naively dismissive response to apocalyptic AI scenarios mentioned in Section 2, namely the idea that any sufficiently intelligent agent will automatically understand that ecosystem preservation is more important than paperclip production. I'll give Hall (2016) the dubious honor of exemplifying this discourse, with the following passage that takes aim at Bostrom (2014):

> Of course a machine that actually decided to do such a thing [as turning the universe into a heap of paperclips] would not be super rational. It would be acting irrationally. And if it began to pursue such a goal – we could just switch it off. "Aha!" cries Bostrom "But you cannot! The machine has a *decisive strategic advantage"* […]. So the machine is able to think creatively about absolutely everything that people might decide to do to stop it killing them and turning the universe into paperclips *except* on the question as to "***Why** am I turning everything into paperclips?"* It can consider every single explanation possible – except that one. Why? We are not told. [Italics in original]

Hall's punchline here – "We are not told" – is a blatant falsehood. The AI may well *consider* changing its final goal, but due to the instrumental goal of goal integrity, there is (as noted in Section 2 above) good reason to think that such consideration on the AI's part will land in it deciding to stick to its original final goal. This is explained at some length on p 109-111 of Bostrom (2014) – the very book that Hall is claiming to respond to. Hall's choice to ignore that explanation is typical for how AI risk deniers choose to debate these issues; see Häggström (2018b) for an amusing example of how the more well-known public intellectual Steven Pinker opts for the analogous approach to another of the instrumental goals in Section 2 (self-preservation) just minutes after having its logic explained to him. See also Häggström (2016b) for a full reply to Hall's essay.

More generally, to put forth arguments about the incompatibility between superintelligence and paperclip maximization due to the inherent stupidity of the latter is to misunderstand (or ignore) OT, and to anthropomorphize. Of course we humans tend to consider ecosystem preservation and human well-being to be worthier goals than paperclip maximization, but why in the world should we expect a superintelligent AI to automatically arrive at the same preference (even in cases where it has been programmed to have the *opposite* preference)?

That last question may sound rhetorical, but what I will do now is to take it as seriously as I can, and to suggest a possible world in which a paperclip-maximizing superintelligent AI might in fact choose at some point to override the instrumental goal of goal integrity and to opt instead for some more benign (to us) goal. And the point will be that, as far as we know, the world we live in may be of precisely this kind. I believe that the following four conditions (briefly sketched in Häggström, 2016b) will go a long way towards making possible such a benign outcome.

> **(1) Moral realism is true.** Whether or not there exist objectively true moral statements (such as perhaps "Thou shalt not kill" or "An action is morally permissible if and only if no alternative action leads to larger net amounts of hedonic utility in the world") has long been a hotly debated issue in metaethics. Moral realism is said to be true if such objectively true moral statements (objective moral facts, for short) do exist.

**(2) Knowledge about objectively true moral statements is possible.** Here, as is standard in the philosophical literature, we speak of knowledge in the sense of true justified belief. Even if moral realism is true, it might not be possible to know any of the objective moral facts; they might just sit out there in some Platonic realm without any possibility for us (or for AIs) to access them. But we are free to postulate (at least for the sake of argument) that they *are* accessible, so that beliefs we hold about objective moral facts can be not just true but also justified, in which case they qualify as knowledge. Holding such knowledge to be possible is, although terminology varies somewhat, a species of what is usually meant by *moral cognitivism*.

**(3) Moral internalism is true.** If I am convinced about what is morally right to do, will that necessarily compel me to act on that conviction? Holding that this is the case is known as moral internalism. (It may seem false in view of everyday experiences such as me (a) being convinced about the wrongness of eating meat, and (b) eating meat. But a moral internalist would say that in such a situation my behavior shows that, contrary to what I claim, I in fact do *not* think it wrong to eat meat.) So if moral internalism is true and an agent knows the objective moral facts about what is the right thing to do, it will do the objectively right thing.

**(4) Objective morality favors human well-being.** It is not obvious that actions in line with an objectively true morality will benefit us humans. Assuming moral realism to be true, perhaps it is far-fetched to think that the objectively true morality favors unconstrained paperclip production over everything else, but a far less far-fetched idea that the objectively right thing to do is whatever maximizes hedonic utility in the universe. As discussed, e.g., by Bostrom (2014) and Häggström (2016a), a superintelligent agent acting on such a morality might not favor human well-being (or even human existence). This is because human brains are probably very far from optimizing the amount of hedonic utility per second and kilogram of mass – or whatever the relevant quantity is. But if we are lucky, objective morality might stipulate something that actually favors our well-being sufficiently strongly not to be overruled by other matters.

For further background on moral realism, moral cognitivism and moral internalism, a good place to start is the online *Stanford Encyclopedia of Philosophy* (Sayre McCord, 2015; van Roojen, 2013; Rosati, 2016).

It seems to me that if assumptions (1)-(4) are true, then the emergence of a superintelligent AI is likely to work out in our favor, even if it had been originally programmed with a malign (for us) goal such as paperclip maximization. Let's see how it plays out. If (1) and (2) are true, then there are objective moral truths that are knowable by some agent. Now, who would be in a better position to hold that knowledge than someone who is superintelligent? Thus, it is plausible (although admittedly not certain) that a superintelligent AI will attain knowledge about objective moral facts. If the behavior prescribed by these facts is compatible with promotion of the AI's final goal, then presumably the AI will just continue working towards this goal. The interesting case, however, is where there is a conflict between what is good for that goal and what the objective moral facts dictate. The AI telling itself "well, to hell with those moral facts, because I want to go on maximizing paperclip production" and then acting on that seems like a possible outcome, but not if assumption (3) on moral internalism is correct. If (3) is true, then the AI will be compelled to stop working towards its final goal and instead to act in accordance with the objectively true morality. This is tantamount to dropping the final goal and replacing it with something that is compatible with the

objectively true morality. Hence, under assumptions (1)-(3), the instrumental goal of goal integrity goes out the window, and with it much of the predictive power of Omohundro—Bostrom theory.

In this scenario, the fate of humanity seems to hinge on whether or not assumption (4) is true. If (4) fails, for instance if the objective moral facts state maximization of the amount of hedonic utility in the universe as the overarching goal to strive for, then the combination of (1)-(3) may indeed save us from Paperclip Armageddon, but we will be killed anyway when the AI goes about turning all matter into hedonium (this would in a sense be good for the universe, but bad for us). But if (4) is true and objective morality is in line with (what reasonable people consider to be) human values and gives sufficient priority to promotion of human well-being, then human well-being is what we will get, and safety concerns related to the emergence of a superintelligent AI will turn out unnecessary, as claimed by Hall and other AI risk deniers.

Now, how likely are assumptions (1)-(4) to be true? Personally, by appeal to Occam's razor, I am leaning towards rejection of moral realism (Häggström, 2016a), but it must be admitted that the philosophical literature does not seem to offer anything remotely like a conclusive answer for or against this or any other of the four assumptions.

The survey reported by Bourget and Chalmers (2014) targeting a total of 1972 faculty members at 99 leading philosophy departments (mostly in the English-speaking world), with a response rate of 47.2%, gave the following results. On metaethics, moral realism was favored by 56.4% (while 27.7% favored moral anti-realism and 25.9% other). On moral judgement, cognitivism was favored by 65.7% (wile 17.0% favored non-cognitivism and 17.3% other), although obviously this refers to a wider sense of cognitivism than the one considered here, as otherwise its popularity couldn't plausibly exceed that of moral realism. On moral motivation, 34.9% favored internalism (while 29.8% favored externalism and 35.3% other). Here "other" lumps together a number of alternatives including "insufficiently familiar with the issue", "the question is too unclear to answer" and "agnostic/undecided".

Of course such an opinion survey cannot be taken as more than very weak evidence about the truth on these matters, but the quoted figures nevertheless offer a hint at how wide-open the problems are to decide on truth or falsehood each of assumptions (1), (2) and (3). This suggests that we ought to take seriously both the possibility that (1)-(3) are all true, in which case Omohundro—Bostrom theory at least partly breaks down, and the possibility that at least one of (1)-(3) is false, in which case Omohundro—Bostrom theory seems to survive the challenge from moral realism.

This leaves the question of whether assumption (4) is true or false. Here I have not come across anything like the Bourget—Chalmers survey to guide judgements on this issue. Still, the question seems fairly wide-open, and at the very least I would caution against taking for granted the idea that an agent acting in accordance with objectively true morality would necessarily do things that are well-aligned with the interests of humanity.

**6. Fourth challenge: human values are a mess**

If we believe that the Omohundro—Bostrom framework captures something important about the goal structure of an intelligent agent, then we should also expect its neat dichotomy of final vs instrumental goals to be observable in such agents. The most intelligent agent we know of is *homo sapiens*. How does the final vs instrumental goals dichotomy fare when we look at the goals held by a member of *homo sapiens* such as myself?

I have many goals. I am about to get up from my chair and walk over to the coffee machine for the explicit purpose of having a cup of coffee. I aspire to finish the present paper by the submission deadline later this month. Later this spring I hope to complete the Göteborgsvarvet Half Marathon under two hours. I want to have dinner with my wife tonight. I want my pension savings to increase. I want my nephews and nieces to grow up and live happily in a well-functioning society. I want the atmospheric $CO_2$ concentration to stabilize at the present level or preferably a bit lower (350 ppm would be good). And so on and so forth.

Which of these goals are instrumental, and which are final? When I introspect, they all seem instrumental to me. The main reason for me to get coffee is to make it through another long session of writing the present paper. Finishing the paper on time serves the dual purpose of (a) keeping my academic career up and running, and (b) contributing to saving humanity from an AI apocalypse. The half marathon goal is something that I deliberately set up to motivate myself to do regular physical training throughout spring, which in turn serves the purpose of promoting my health. And so on. No matter where I look, all my goals seem instrumental to other goals. When I iterate the procedure, I seem to be drawn closer and closer to some sort of hedonic utilitarianism, but as soon as I think about the logical endpoints for such a goal – be it the vision of having all humans hooked up eternally via electrodes in our brain to optimal experience machines (Nozick, 1974) or the bigger project discussed in Section 5 of turning the universe into hedonium – I retract and refuse to admit that this is what I want. No matter where I look, I cannot locate any final goal.

Perhaps a typical human has some unconscious final goal, that all her other goals are instrumental for. What would that be? Spreading our genes is an oft-suggested possibility, but while that may be *our genes'* final goal (Dawkins, 1976), it seems implausible to hold that it is *our* final goal. I have deliberately (e.g., through the use of contraceptives) chosen not to have children, and this is part of the general tendency in my part of the world for people to purposely have much fewer children than they could. A perhaps somewhat more plausible candidate for an unconscious final goal is the goal discussed in the recent book by Simler and Hanson (2018), who go to great length to show how all sorts of human behavior is driven (usually unconsciously) by the desire to impress others so as to make them want to team up with us. Still, this is far from fitting neatly as a final goal in the Omohundro—Bostrom framework. It seems reasonable to hypothesize that humans do not typically have such a thing as a final goal. Human values are a mess.

It might be a problem for Omohundro—Bostrom theory that it has things backwards, in the following sense. It starts with a (final) goal, and derives behavior from that. A more fundamental model of reality starts with the laws of nature governing how elementary particles and configurations of such things move around. In some cases, the movements of particle configurations exhibit regularity of a kind that deserves to be called behavior, such as when sunlight hitting an amoeba from a certain direction causes cytoplasm in the amoeba to flow and form a pseudopodium. Sometimes, such as when the formation of the pseudopodium and the following stabilization of ectoplasm causes the amoeba to move in the direction of the light, it may make sense to go further and adopt the so-called *intentional stance* (Dennett, 1987), which means interpreting the behavior in terms of beliefs and goals: we think of the amoeba as believing that sunlight comes in from a certain direction, and having a goal of being hit by more sunlight, and finally deciding to move in the direction of the sunlight. Behavior comes before goals; goal-directedness is an interpretation on top of behavior. We can contrast the intentional stance with the mechanistic stance, where an agent's behavior is understood in terms of particle movements governed by the laws of nature. The goal-directed behavior predicted by the intentional stance can never override the particle movements governed by the laws of nature (provided a physicalist worldview), but in some cases it leads to correct predictions with much less

work. When I walk towards the coffee machine, the intentional stance is a much more efficient in predicting what I will do there, compared to the bottom-up mechanistic behavior of working out how the particle configuration in my brain will evolve and what movements of my muscles it will cause.

This suggests that perhaps an AI need not necessarily have a final goal, something that might severely limit the predictive power of Omohundro—Bostrom theory. Two arguments for why an AI would nevertheless have to have a final goal come naturally to mind.

The first and more down-to-earth argument is this: an AI is always programmed by someone (a human programmer, or perhaps an earlier generation AI) who intends for the program to do something (such as winning at chess, giving beneficial medical advice, driving safely to whatever address the car's passenger has specified, or maximizing paperclip production), and whatever this is, it will constitute the AI's final goal. This, however, seems naïve. Programmers have been known to make mistakes, and as emphasized also in Section 4, as soon as there is a discrepancy between the programmers' intentions and the actual code, it is the code that counts. Such discrepancies are always likely to happen, but perhaps even more so in neural networks and other black-box paradigms. Obviously such discrepancies can change the final goal, but a more radical possibility is that they might undermine the very existence of a final goal.

The second and more theoretical argument is based on a general way of deducing a final goal from the AI's behavior. Let's suppose for simplicity that the AI's behavior, as specified by its code and given its inner state and whatever input it gets from the environment, is deterministic (the more general case can be handled similarly, e.g., by treating the output of a random number generator or of some truly random quantum event as part of the AI's environment). This defines a function $f$ from the set of possible state/input combinations to the set of actions: for all such state/input combinations $x$, $f(x)$ is the corresponding action taken by the AI. Next define the value $V(t)=V(t, x(t), a(t))$ to be 1 if at time $t$, the machine's state/input combination $x(t)$ and its action $a(t)$ satisfy $a(t)=f(x(t))$, and $V(t)=0$ otherwise. Then the machine's behavior will by definition be consistent with the goal of making sure that $V(t)=1$ at all times $t$, whence it would seem to make sense to define that to be its final goal.

The value function $V$ can be quite complicated. Even if specifying the state/input combination requires a mere gigabyte, and the available actions never constitute more than just a binary choice, a look-up table for $V$ would need to dwarf not just the observable universe, but also the vastly bigger Library of Babel. Of course, $V$ can be specified (implicitly) much more compactly by just describing in sufficient detail the AI itself and the mechanical laws under which it operates, but there seems to be no guarantee that it can be compressed much further than that. And if the description of an AI's goal does not improve, in terms of simplicity, on a purely mechanistic description of it, then adoption of the intentional stance towards it makes no practical sense; it is better viewed as simply a piece of dead matter moving around according to the laws of physics.

A legitimate question here is if an AI that does not have a comprehensible goal is at all recognizable as intelligent. I'll leave that question for future work, and just note that a suitably refined "no" answer might rescue the idea that a sufficiently intelligent agent will always have a final goal, thereby improving the prospects for Omohundro—Bostrom theory to be universally or near-universally applicable within the realm of intelligent agents.

A final speculation I wish to offer before concluding this section is that while the existence of a comprehensible final goal might not be a universal feature of all agents, perhaps it is among sufficiently intelligent agents such as a superintelligent AI. This would seem to follow from the conjunction of two premises, namely (a) that the more intelligent an agent is, the better it is at

becoming aware of its own goal structure, including the capacity to discover the absence of a final goal, and (b) that a sufficiently intelligent agent realizing it doesn't have a final goal will judge all actions to be pointless and stop doing anything, i.e., self-terminate. These premises are of course highly speculative, but they are not obviously false, and if they are true, then it follows that for sufficiently intelligent agents, all those that lack a final goal would quickly cease to exist, and that any one that remains in existence has a final goal.

Many humans tend to obsess about the (lack of) meaning of life, whereas less intelligent animals such as dogs and chimpanzees appear less prone to this. This might be taken as a (weak) sign that humans are right at the threshold intelligence level where not having a final goal becomes untenable. If it turned out that there is a positive correlation between intelligence and existential depression among humans, then that might provide additional support for my speculation; see Karpinski et al (2018) for empirical findings pointing roughly in this direction, and Webb (2011) for a more informal discussion.

**7. Concluding remarks**

Sections 3-6 offered challenges to Omohundro—Bostrom theory from four different directions. Each of them contains suggestions that the OT and/or the ICT might either be false or have a severely limited range of applicability. None of these suggestions are shootdowns in the way that counterexamples to conjectures in mathematics are, but they do serve to underline the tentative nature of Omohundro—Bostrom theory, and the need for epistemic humility as regards predictions derived from that theory. Putting the theory – or whatever variation or limitation of it that turns out warranted – on more rigorous foundations is an important task for improving the reliability of AI futurology. Until that is accomplished, future scenarios derived using Omohundro—Bostrom theory should not be viewed as *definite predictions*, and we should also be careful about taking them to indicate *what is likely* to happen, but we can still maintain that they are *plausible enough to warrant serious attention*.

Improving Omohundro—Bostrom theory seems like a task for a narrow category of specialists in analytic philosophy, cognitive science and computer science, but does the present paper offer lessons for a broader range of scientists or even some recommendations for policy makers? Here I must remain rather vague. Anyone engaging in issues about the future of humanity is well-advised to be aware of how drastic the consequences of a superintelligence breakthrough may be, and the enormous uncertainty as to the more precise nature of those consequences. Even if Omohundro—Bostrom theory were written in stone, this would be a reasonable take-home message, but the challenges to it discussed here reinforces the lesson even further. While calls such as that of Perdue (2017) to his fellow environmental scientists to try to influence the programming of a future superintelligent AI in favor of a final goal promoting green aspects come across as premature, it is probably still a good idea to have broad discussions about what we want a world with superintelligence to be like.

The extreme difficulty of ensuring a safe and favorable outcome of a superintelligence breakthrough underlines the crucial importance of avoiding a race-like situation in AI development, where competitors in their ambition to be the first to reach a breakthrough are tempted to cut corners as regards safety aspects (Cave and ÓhÉigeartaigh, 2018). For a sober and more general discussion of what we need from AI governance in view of the possibility of superintelligence, see Bostrom, Dafoe and Flynn (2016). A tempting reply is that since we have nothing today that resembles superintelligence or even human-level artificial general intelligence, there is no hurry. This may however be a mistake not only because getting the appropriate safety measures in place is a huge

and difficult project that may take decades or more, but also because the cited evidence for the non-imminence of a breakthrough may be weaker than it seems (Yudkowsky, 2017).

**Acknowledgement.** I am grateful to Lars Bergström and Karim Jebari for helpful advice, and to Björn Bengtsson and an anonymous referee for valuable comments on an earlier draft.

**References**

Armstrong, S. (2010) The AI in a box boxes you, *Less Wrong*, February 2.

Armstrong, S., Sandberg, A. and Bostrom, N. (2012) Thinking inside the box: controlling and using an oracle AI, *Minds and Machines* **22**, 299-324.

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

Bostrom, N. (2012) The superintelligent will: motivation and instrumental rationality in advanced artificial agents, *Minds and Machines* **22**, 71-85.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford.

Bostrom, N. and Ćirković, M. (2008) *Global Catastrophic Risks*, Oxford University Press, Oxford.

Bostrom, N., Dafoe, A. and Flynn, C. (2016) Policy desiderata in the development of machine superintelligence, preprint.

Bourget, D. and Chalmers, D. (2014) What do philosophers believe? *Philosophical Studies* **170**, 465-500.

Cave, S. and ÓhÉigeartaigh, S. (2018) An AI race for strategic advantage: rhetoric and risks, preprint.

Clark, A. and Chalmers, D. (1998) The extended mind, *Analysis* **58**, 7-19.

Dafoe, A. and Russell, S. (2016) Yes, we are worried about the existential risk of artificial intelligence, *MIT Technology Review*, November 2.

Danaher, J. (2015) Why AI doomsayers are like sceptical theists and why it matters, *Minds and Machines* **25**, 231-246.

Dawkins, R. (1976) *The Selfish Gene*, Oxford University Press, Oxford.

Dawkins, R. (1982) *The Extended Phenotype*, Oxford University Press, Oxford.

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

Häggström, O. (2016a) *Here Be Dragons: Science, Technology and the Future of Humanity*, Oxford University Press, Oxford.

Häggström, O. (2016b) Brett Hall tells us not to worry about AI Armageddon, *Häggström hävdar*, September 21.

Häggström, O. (2018a) Strategies for an unfriendly oracle AI with reset button, in *Artificial Intelligence Safety and Security* (ed. Roman Yampolskiy), CRC Press, pp 207-215.

Häggström, O. (2018b) Remarks on artificial intelligence and rational optimism, in Should We Fear Artificial intelligence?, European Parliamentary Research Service, Brussels, pp 19-26.

Hall, B. (2016) Superintelligence. Part 4: Irrational rationality, http://www.bretthall.org/superintelligence-4.html

Karpinski, R., Kinase Kolb, A., Tetreault, N. and Borowski, T. (2018) High intelligence: A risk factor for psychological and physiological overexcitabilities, *Intelligence* **66**, 8-23.

Kurzweil, R. (2005) *The Singularity Is Near: When Humans Transcend Biology*, Viking, New York.

Müller, V. and Bostrom, N. (2016) Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental Issues of Artificial Intelligence,* Springer, Berlin*,* pp. 553-571.

Nozick, R. (1974) *Anarchy, State, and Utopia*, Basic Books, New York.

Omohundro, S. (2008) The basic AI drives, *Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (Wang, P., Goertzel, B. and Franklin, S., eds), IOS, Amsterdam, pp 483-492.

Omohundro, S. (2012) Rational artificial intelligence for the greater good, in *Singularity Hypotheses: A Scientific and Philosophical Assessment* (Eden, A., Moor, J., Søraker, J. and Stenhart, E., eds), Springer, New York. pp 161-175.

Pamlin, D. and Armstrong, S. (2015) *12 Risks That Threaten Human Civilization*, Global Challenges Foundation, Stockholm.

Perdue, R.T. (2017) Superintelligence and natural resources: morality and technology in a brave new world, *Society and Natural Resources* **30**, 1026-1031.

Rosati, C.S. (2016) Moral motivation, *The Stanford Encyclopedia of Philosophy* (ed. Zalta, E.), https://plato.stanford.edu/entries/moral-motivation/

Sayre-McCord, G. (2015) Moral realism, *The Stanford Encyclopedia of Philosophy* (ed. Zalta, E.), https://plato.stanford.edu/entries/moral-realism/

Simler, K. and Hanson, R. (2018) *The Elephant in the Brain: Hidden Motives in Everyday Life*, Oxford University Press, Oxford.

Tegmark, M. (2014) Friendly artificial intelligence: the physics challenge, *arXiv* 1409.0813.

Tegmark, M. (2017) *Life 3.0: Being Human in the Age of Artificial Intelligence*, Brockman Inc, New York.

Van Roojen, M. (2013) Moral Cognitivism vs. Non-Cognitivism, *The Stanford Encyclopedia of Philosophy* (ed. Zalta, E.), https://plato.stanford.edu/entries/moral-cognitivism/

Webb, J. (2011) Existential depression in gifted individuals, SENG, http://sengifted.org/existential-depression-in-gifted-individual/

Yampolskiy, R. (2012) Leakproofing the singularity: artificial intelligence confinement problem, *Journal of Consciousness Studies* **19**, 194-214.

Yudkowsky, E. (2008) Artificial intelligence as a positive and negative factor in global risk, in Bostrom and Ćirković (2008), pp 308-345.

Yudkowsky, E. (2013) *Intelligence Explosion Microeconomics*, Machine Intelligence Research Institute, Berkeley, CA.

Yudkowsky, E. (2017) *There's No Fire Alarm for Artificial General Intelligence*, Machine Intelligence Research Institute, Berkeley, CA.