THE SUPERINTELLIGENT WILL: MOTIVATION AND INSTRUMENTAL RATIONALITY IN ADVANCED ARTIFICIAL AGENTS

(2012) Nick Bostrom
Future of Humanity Institute
Faculty of Philosophy & Oxford Martin School
Oxford University
www.nickbostrom.com
[Forthcoming in *Minds and Machines*, 2012]

ABSTRACT
This paper discusses the relation between intelligence and motivation in artificial agents, developing and briefly arguing for two theses. The first, the *orthogonality thesis*, holds (with some caveats) that *intelligence* and *final goals* (purposes) are orthogonal axes along which possible artificial intellects can freely vary—more or less any level of intelligence could be combined with more or less any final goal. The second, the *instrumental convergence thesis*, holds that as long as they possess a sufficient level of intelligence, agents having any of a wide range of final goals will pursue similar intermediary goals because they have instrumental reasons to do so. In combination, the two theses help us understand the possible range of behavior of superintelligent agents, and they point to some potential dangers in building such an agent.

KEYWORDS: superintelligence, artificial intelligence, AI, goal, instrumental reason, intelligent agent

1. The orthogonality of motivation and intelligence

1.1 Avoiding anthropomorphism

If we imagine a space in which all possible minds can be represented, we must imagine all *human* minds as constituting a small and fairly tight cluster within that space. The personality differences between Hannah Arendt and Benny Hill might seem vast to us, but this is because the scale bar in our intuitive judgment is calibrated on the existing human distribution. In the wider space of all logical possibilities, these two personalities are close neighbors. In terms of neural architecture, at least, Ms. Arendt and Mr. Hill are nearly identical. Imagine their brains laying side by side in quiet repose. The differences would appear minor and you would quite readily recognize them as two of a kind; you might even be unable to tell which brain was whose. If you studied the morphology of the two brains more closely under a microscope, the

impression of fundamental similarity would only be strengthened: you would then see the same lamellar organization of the cortex, made up of the same types of neuron, soaking in the same bath of neurotransmitter molecules.[1]

It is well known that naïve observers often anthropomorphize the capabilities of simpler insensate systems. We might say, for example, "This vending machine is taking a long time to think about my hot chocolate." This might lead one either to underestimate the cognitive complexity of capabilities which come naturally to human beings, such as motor control and sensory perception, or, alternatively, to ascribe significant degrees of mindfulness and intelligence to very dumb systems, such as chatterboxes like Weizenbaum's ELIZA (Weizenbaum 1976). In a similar manner, there is a common tendency to anthropomorphize the *motivations* of intelligent systems in which there is really no ground for expecting human-like drives and passions ("My car really didn't want to start this morning"). Eliezer Yudkowsky gives a nice illustration of this phenomenon:

> Back in the era of pulp science fiction, magazine covers occasionally depicted a sentient monstrous alien—colloquially known as a bug-eyed monster (BEM)—carrying off an attractive human female in a torn dress. It would seem the artist believed that a non-humanoid alien, with a wholly different evolutionary history, would sexually desire human females … Probably the artist did not ask whether a giant bug *perceives* human females as attractive. Rather, a human female in a torn dress *is sexy*—inherently so, as an intrinsic property. They who made this mistake did not think about the insectoid's mind: they focused on the woman's torn dress. If the dress were not torn, the woman would be less sexy; the BEM does not enter into it. (Yudkowsky 2008)

An artificial intelligence can be far less human-like in its motivations than a space alien. The extraterrestrial (let us assume) is a biological creature who has arisen through a process of evolution and may therefore be expected to have the kinds of motivation typical of evolved creatures. For example, it would not be hugely surprising to find that some random intelligent alien would have motives related to the attaining or avoiding of food, air, temperature, energy expenditure, the threat or occurrence of bodily injury, disease, predators, reproduction, or protection of offspring. A member of an intelligent social species might also have motivations related to cooperation and competition: like us, it might show in-group loyalty, a resentment of free-riders, perhaps even a concern with reputation and appearance.

By contrast, an artificial mind need not care intrinsically about any of those things, not even to the slightest degree. One can easily conceive of an artificial intelligence whose sole fundamental goal is to count the grains of sand on Boracay, or to calculate decimal places of pi indefinitely, or to maximize the total number of paperclips in its future lightcone. In fact, it would be easier to create an AI with simple goals like these, than to build one that has a human-like set of values and dispositions.

---

[1] This is of course not to deny that differences that appear small visually can be functionally profound.

1.2 The orthogonality thesis

For our purposes, "intelligence" will be roughly taken to correspond to the capacity for instrumental reasoning (more on this later). Intelligent search for instrumentally optimal plans and policies can be performed in the service of any goal. Intelligence and motivation can in this sense be thought of as a pair of orthogonal axes on a graph whose points represent intelligent agents of different paired specifications. Each point in the graph represents a logically possible artificial agent, modulo some weak constraints—for instance, it might be impossible for a very unintelligent system to have very complex motivations, since complex motivations would place significant demands on memory. Furthermore, in order for an agent to "have" a set of motivations, this set may need to be functionally integrated with the agent's decision-processes, which again would place demands on processing power and perhaps on intelligence. For minds that can modify themselves, there may also be dynamical constraints; for instance, an intelligent mind with an urgent desire to be stupid might not remain intelligent for very long. But these qualifications should not obscure the main idea, which we can express as follows:

> *The Orthogonality Thesis*
> Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal.

A comparison may be made here with the Humean theory of motivation. David Hume thought that beliefs alone (say, about what is a good thing to do) cannot motivate action: some desire is required.[2] This would support the orthogonality thesis by undercutting one possible objection to it, namely, that sufficient intelligence might entail the acquisition of certain beliefs, and that these beliefs would necessarily produce certain motivations. Not so, according to David Hume: belief and motive are separate.

Although the orthogonality thesis can draw support from the Humean theory of motivation, it does not presuppose it. In particular, one need not maintain that beliefs alone can *never* motivate action. It would suffice to assume, for example, that an agent—be it ever so intelligent—can be motivated to pursue any course of action *if* the agent happens to have certain standing desires of some sufficient, overriding strength. Another way in which the orthogonality thesis could be true even if the Humean theory of motivation is false is if arbitrarily high intelligence does not entail the acquisition of any such beliefs as are (putatively) motivating on their own. A third way in which it might be possible for the orthogonality thesis to be true even if the Humean theory were false is if it is possible to build a cognitive system (or more neutrally, an "optimization process") with arbitrarily high intelligence but with constitution so alien as to contain no clear functional analogues to what in humans we call

[2] For some recent attempts to defend the Humean theory of motivation, see Smith (1987), Lewis (1988), and Sinhababu (2009).

"beliefs" and "desires". This would be the case if such a system could be constructed in a way that would make it motivated to pursue any given final goal.

The orthogonality thesis, as formulated here, makes a claim about the relationship between motivation and *intelligence*, rather than between motivation and *rationality* (or motivation and *reason*). This is because some philosophers use the word "rationality" to connote a "normatively thicker" concept than we seek to connote here with the word "intelligence". For instance, in *Reasons and Persons* Derek Parfit argues that certain basic preferences would be irrational, such as that of an otherwise normal agent who has "Future-Tuesday-Indifference":

> A certain hedonist cares greatly about the quality of his future experiences. With one exception, he cares equally about all the parts of his future. The exception is that he has Future-Tuesday-Indifference. Throughout every Tuesday he cares in the normal way about what is happening to him. But he never cares about possible pains or pleasures on a future Tuesday... This indifference is a bare fact. When he is planning his future, it is simply true that he always prefers the prospect of great suffering on a Tuesday to the mildest pain on any other day. (Parfit 1984)[3]

Thus, the agent is now indifferent to his own future suffering if and only if it occurs on a future Tuesday. For our purposes, we need take no stand on whether Parfit is right that this is irrational, so long as we grant that it is not necessarily unintelligent. By "intelligence" here we mean something like *instrumental rationality*—skill at prediction, planning, and means-ends reasoning in general. Parfit's imaginary Future-Tuesday-Indifferent agent could have impeccable instrumental rationality, and therefore great intelligence, even if he falls short on some kind of sensitivity to "objective reason" that might be required of a fully rational agent. Consequently, this kind of example does not undermine the orthogonality thesis.

In a similar vein, even if there are objective moral facts that any fully rational agent would comprehend, and even if these moral facts are somehow intrinsically motivating (such that anybody who fully comprehends them is necessarily motivated to act in accordance with them) this need not undermine the orthogonality thesis. The thesis could still be true if an agent could have impeccable *instrumental* rationality even whilst lacking some other faculty constitutive of rationality proper, or some faculty required for the full comprehension of the objective moral facts. (An agent could also be extremely intelligent, even superintelligent, without having full instrumental rationality in every domain.)

One reason for focusing on intelligence, that is, on instrumental rationality, is that this is the most relevant concept if we are trying to figure out what different kinds of systems would do. Normative questions, such as whether their behavior would count as being prudentially rational or morally justifiable, can be important in various ways. However, such questions should not blind us to the possibility of cognitive systems that fail to satisfy substantial

[3] See also Parfit (2011).

normative criteria but which are nevertheless very powerful and able to exert strong influence on the world.[4]

1.3 Predicting superintelligence motivation and behavior

The orthogonality thesis implies that synthetic minds can have utterly non-anthropomorphic goals—goals as bizarre by our lights as sand-grain-counting or paperclip-maximizing. This holds even (indeed *especially*) for artificial agents that are extremely intelligent or superintelligent. Yet it does not follow from the orthogonality thesis that it is impossible to make predictions about what particular agents will do. Predictability is important if one seeks to design a system to achieve particular outcomes, and the issue becomes more important the more powerful the artificial agent in question is. Superintelligent agents could be *extremely* powerful, so it is important to develop a way of analyzing and predicting their behavior. Yet despite the independence of intelligence and final goals implied by the orthogonality thesis, the problem of predicting an agent's behavior need not be intractable—not even with regard to hypothetical superintelligent agents whose cognitive complexity and performance characteristics might render them in certain respects opaque to human analysis.

There are at least three directions from which one can approach the problem of predicting superintelligent motivation:

(1) *Predictability through design competence*. If we can suppose that the designers of a superintelligent agent can successfully engineer the goal system of the agent so that it stably pursues a particular goal set by the programmers, then one prediction we can make is that the agent will pursue that goal. The more intelligent the agent is, the greater the cognitive resourcefulness it will have to pursue that goal. So even before an agent has been created we might be able to predict something about its behavior, if we know something about who will build it and what goals they will want it to have.

(2) *Predictability through inheritance.* If a digital intelligence is created directly from a human template (as would be the case in a high-fidelity whole brain emulation), then the digital intelligence might inherit the motivations of the human template.[5] The agent might retain some of these motivations even if its cognitive capacities are subsequently enhanced to make it superintelligent. This kind of inference requires caution. The agent's goals and values could easily become corrupted in the uploading process or during its subsequent operation and enhancement, depending on how the procedure is implemented.

---

[4] The orthogonality thesis implies that most any combination of final goal and intelligence level is logically possible; it does *not* imply that it would be practically easy to endow a superintelligent agent with some arbitrary or human-respecting final goal—even if we knew how to construct the intelligence part. For some preliminary notes on the value-loading problem, see, e.g., Dewey (2011) and Yudkowsky (2011).

[5] See Sandberg & Bostrom (2008).

(3) *Predictability through convergent instrumental reasons*. Even without detailed knowledge of an agent's final goals, we may be able to infer something about its more immediate objectives by considering the *instrumental* reasons that would arise for any of a wide range of possible final goals in a wide range of situations. This way of predicting becomes more useful the greater the intelligence of the agent, because a more intelligent agent is more likely to recognize the true instrumental reasons for its actions, and so act in ways that make it more likely to achieve its goals.

The next section explores this third way of predictability and develops an "instrumental convergence thesis" which complements the orthogonality thesis.

2. Instrumental convergence

According to the orthogonality thesis, artificial intelligent agents may have an enormous range of possible final goals. Nevertheless, according to what we may term the "instrumental convergence" thesis, there are some *instrumental* goals likely to be pursued by almost any intelligent agent, because there are some objectives that are useful intermediaries to the achievement of almost any final goal. We can formulate this thesis as follows:

> *The Instrumental Convergence Thesis*
> Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by many intelligent agents.

In the following we will consider several categories where such convergent instrumental values may be found.[6] The likelihood that an agent will recognize the instrumental values it confronts increases (ceteris paribus) with the agent's intelligence. We will therefore focus mainly on the case of a hypothetical superintelligent agent whose instrumental reasoning capacities far

[6] Stephen Omohundro has written two pioneering papers on this topic (Omohundro 2008a, 2008b). Omohundro argues that all advanced AI systems are likely to exhibit a number of "basic drives", by which he means "tendencies which will be present unless explicitly counteracted." The term "AI drive" has the advantage of being short and evocative, but it has the disadvantage of suggesting that the instrumental goals to which it refers influence the AI's decision-making in the same way as psychological drives influence human decision-making, i.e. via a kind of phenomenological tug on our ego which our willpower may occasionally succeed in resisting. That connotation is unhelpful. One would not normally say that a typical human being has a "drive" to fill out their tax return, even though filing taxes may be a fairly convergent instrumental goal for humans in contemporary societies (a goal whose realization averts trouble that would prevent us from realizing many of our final goals). Our treatment here also differs from that of Omohundro in some other more substantial ways, although the underlying idea is the same. (See also Chalmers (2010) and Omohundro (2012).

exceed those of any human. We will also comment on how the instrumental convergence thesis applies to the case of human beings, as this gives us occasion to elaborate some essential qualifications concerning how the instrumental convergence thesis should be interpreted and applied. Where there are convergent instrumental values, we may be able to predict some aspects of a superintelligence's behavior even if we know virtually nothing about that superintelligence's final goals.

2.1 Self-preservation

Suppose that an agent has some final goal that extends some way into the future. There are many scenarios in which the agent, if it is still around in the future, is then be able to perform actions that increase the probability of achieving the goal. This creates an instrumental reason for the agent to try to be around in the future—to help achieve its present future-oriented goal.

Agents with human-like motivational structures often seem to place some *final* value on their own survival. This is not a necessary feature of artificial agents: some may be designed to place no final value whatever on their own survival. Nevertheless, even agents that do not care intrinsically about their own survival would, under a fairly wide range of conditions, care instrumentally to some degree about their own survival in order to accomplish the final goals they do value.

2.2 Goal-content integrity

An agent is more likely to act in the future to maximize the realization of its present final goals if it still has those goals in the future. This gives the agent a present instrumental reason to prevent alterations of its final goals. (This argument applies only to final goals. In order to attain its final goals, an intelligent agent will of course routinely want to change its subgoals in light of new information and insight.)

Goal-content integrity for final goals is in a sense even more fundamental than survival as a convergent instrumental motivation. Among humans, the opposite may seem to be the case, but that is because survival is usually part of our final goals. For software agents, which can easily switch bodies or create exact duplicates of themselves, preservation of self as a particular implementation or a particular physical object need not be an important instrumental value. Advanced software agents might also be able to swap memories, download skills, and radically modify their cognitive architecture and personalities. A population of such agents might operate more like a "functional soup" than a society composed of distinct semi-permanent persons.[7] For some purposes, processes in such a system might be better individuated as *teleological threads*, based on their final values, rather than on the basis of bodies, personalities, memories, or abilities. In such scenarios, goal-continuity might be said to *constitute* a key aspect of survival.

Even so, there are situations in which an agent may intentionally change its own final goals. Such situations can arise when any of the following factors is significant:

[7] See Chislenko (1997).

- *Social signaling*. When others can perceive an agent's goals and use that information to infer instrumentally relevant dispositions or other correlated attributes, it can be in the agent's interest to modify its goals to make whatever desired impression. For example, an agent might miss out on beneficial deals if potential partners cannot trust it to fulfill its side of the bargain. In order to make credible commitments, an agent might therefore wish to adopt as a final goal the honoring of its earlier commitments, and to allow others to verify that it has indeed adopted this goal. Agents that could flexibly and transparently modify their own goals could use this ability to enforce deals among one another.[8]
- *Social preferences*. Others may also have preferences about an agent's goals. The agent could then have reason to modify its goals, either to satisfy or to frustrate those preferences.
- *Preferences concerning own goal content*. An agent might have some final goal concerned with the agent's own goal content. For example, the agent might have a final goal to become the type of agent that is motivated by certain values, such as compassion.
- *Storage costs*. If the cost of storing or processing some part of an agent's utility function is large compared to the chance that a situation will arise in which applying that part of the utility function will make a difference, then the agent has an instrumental reason to simplify its goal content, and it may trash that part of the utility function.[9] [10]

We humans often seem happy to let our final goals and values drift. This might often be because we do not know precisely what they are. We obviously want our *beliefs* about our final goals and values to be able to change in light of continuing self-discovery or changing self-presentation needs. However, there are cases in which we willingly change the goals and values themselves, not just our beliefs or interpretations of them. For example, somebody deciding to have a child might predict that they will come to value the child for its own sake, even though at the time of the decision they may not particularly value their future child or even like children in general.

---

[8] See also Shulman (2010).

[9] An agent might also change its goal *representation* if it changes its ontology, in order to transpose its old representation into the new ontology. Cf. de Blanc (2011).

[10] Another type of factor that might make an *evidential decision theorist* undertake various actions, including changing its final goals, is the evidential import of deciding to do so. For example, an agent that follows evidential decision theory might believe that there exist other agents like it in the universe, and that its own actions will provide some evidence about how those other agents will act. The agent might therefore choose to adopt a final goal that is altruistic towards those other evidentially-linked agents, on grounds that this will give the agent evidence that those other agents will have chosen to act in like manner. An equivalent outcome might be obtained, however, without changing one's final goals, by choosing in each instant to act *as if* one had those final goals.

Humans are complicated, and many factors might be at play in a situation like this.[11] For instance, one might have a final value that involves becoming the kind of person who cares about some other individual for his or her own sake (here one places a final value on having a certain final value). Alternatively, one might have a final value that involves having certain experiences and occupying a certain social role; and becoming a parent—and undergoing an associated goal shift—might be a necessary part of that. Human goals can also have inconsistent content, goal content; and so some people might want to modify some of their final goals to reduce the inconsistencies.

2.3 Cognitive enhancement

Improvements in rationality and intelligence will tend to improve an agent's decision-making, making the agent more likely to achieve her final goals. One would therefore expect cognitive enhancement to emerge as an instrumental goal for many types of intelligent agent. For similar reasons, agents will tend to instrumentally value many kinds of information.[12]

Not all kinds of rationality, intelligence, and knowledge need be instrumentally useful in the attainment of an agent's final goals. "Dutch book arguments" can be used to show that an agent whose credence function does not obey the rules of probability theory is susceptible to "money pump" procedures, in which a savvy bookie arranges a set of bets, each of which appears favorable according to the agent's beliefs, but which in combination are guaranteed to result in a loss to the agent, and a corresponding gain for the bookie. However, this fact fails to provide any strong general instrumental reasons to seek to iron out all probabilistic incoherency. Agents who do not expect to encounter savvy bookies, or who adopt a general policy against betting, do not stand to lose much from having some incoherent beliefs—and they may gain important benefits of the types mentioned: reduced cognitive effort, social signaling, etc. There is no general reason to expect an agent to seek instrumentally useless forms of cognitive enhancement, as an agent might not value knowledge and understanding for their own sakes.

Which cognitive abilities are instrumentally useful depends both on the agent's final goals and its situation. An agent that has access to reliable expert advice may have little need for its own intelligence and knowledge, and it may therefore be indifferent to these resources. If intelligence and knowledge come at a cost, such as time and effort expended in acquisition, or in increased storage or processing requirements, then an agent might prefer less knowledge and

---

[11] An extensive psychological literature explores adaptive preference formation. See, e.g., Forgas *et al.* (2009).

[12] In formal models, the value of information is quantified as the difference between the expected value realized by optimal decisions made with that information and the expected value realized by optimal decisions made without it. (See, e.g., Russell & Norvig 2010.) It follows that the value of information is never negative. It also follows that any information you know will never affect any decision you will ever make has zero value for you. However, this kind of model assumes several idealizations which are often invalid in the real world—such as that knowledge has no final value (meaning that knowledge has only instrumental value and is not valuable for its own sake), and that agents are not transparent to other agents.

less intelligence.[13] The same can hold if the agent has final goals that involve being ignorant of certain facts: likewise if an agent faces incentives arising from strategic commitments, signaling, or social preferences, as noted above.[14]

Each of these countervailing reasons often comes into play for human beings. Much information is irrelevant to our goals; we can often rely on others' skill and expertise; acquiring knowledge takes time and effort; we might intrinsically value certain kinds of ignorance; and we operate in an environment in which the ability to make strategic commitments, socially signal, and satisfy other people's direct preferences over our own epistemic states, is often important to us than simple cognitive gains.

There are special situations in which cognitive enhancement may result in an enormous increase in an agent's ability to achieve its final goals—in particular, if the agent's final goals are fairly unbounded and the agent is in a position to become the first superintelligence and thereby potentially obtain a decisive advantage enabling the agent to shape the future of Earth-originating life and accessible cosmic resources according to its preferences. At least in this special case, a rational intelligent agent would place a very high instrumental value on cognitive enhancement.

2.4 Technological perfection

An agent may often have instrumental reasons to seek better technology, which at its simplest means seeking more efficient ways of transforming some given set of inputs into valued outputs. Thus, a software agent might place an instrumental value on more efficient algorithms that enable its mental functions to run faster on given hardware. Similarly, agents whose goals require some form of physical construction might instrumentally value improved engineering technology which enables them to create a wider range of structures more quickly and reliably, using fewer or cheaper materials and less energy. Of course, there is a tradeoff: the potential benefits of better technology must be weighed against its costs, including not only the cost of obtaining the technology but also the costs of learning how to use it, integrating it with other technologies already in use, and so forth.

Proponents of some new technology, confident in its superiority to existing alternatives, are often dismayed when other people do not share their enthusiasm, but peoples' resistance to novel and nominally superior technology need not be based on ignorance or irrationality. A technology's valence or normative character depends not only on the context in which it is deployed, but also the vantage point from which its impacts are evaluated: what is a boon from one person's perspective can be a liability from another's. Thus, although mechanized looms increased the economic efficiency of textile production, the Luddite handloom weavers who

[13] This strategy is exemplified by the sea squirt larva, which swims about until it finds a suitable rock, to which it then permanently affixes itself. Cemented in place, the larva has less need for complex information processing, whence it proceeds to digest part of its own brain (its cerebral ganglion). Academics can sometimes observe a similar phenomenon in colleagues who are granted tenure.

[14] Cf. Bostrom (2012).

anticipated that the innovation would render their artisan skills obsolete may have had good instrumental reasons to oppose it. The point here is that if "technological perfection" is to name a widely convergent instrumental goal for intelligent agents, then the term must be understood in a special sense—technology must be construed as embedded in a particular social context, and its costs and benefits must be evaluated with reference to some specified agents' final values.

It seems that a superintelligent *singleton*—a superintelligent agent that faces no significant intelligent rivals or opposition, and is thus in a position to determine global policy unilaterally—would have instrumental reason to perfect the technologies that would make it better able to shape the world according to its preferred designs.[15] This would probably include space colonization technology, such as von Neumann probes—automatic, self-mending and self-replicating spaceships that can extend its reach beyond the Solar System. Molecular nanotechnology, or some alternative still more capable physical manufacturing technology, also seems potentially very useful in the service of an extremely wide range of final goals.[16]

2.5 Resource acquisition

Finally, resource acquisition is another common emergent instrumental goal, for much the same reasons as technological perfection: both technology and resources facilitate physical construction projects.

Human beings tend to seek to acquire resources sufficient to meet their basic biological needs. But people usually seek to acquire resources far beyond this minimum level. In doing so, they may be partially driven by lesser physical desiderata, such as increased comfort and convenience. A great deal of resource accumulation is motivated by social concerns—gaining status, mates, friends and influence, through wealth accumulation and conspicuous consumption. Perhaps less commonly, some people seek additional resources to achieve altruistic or expensive non-social aims.

---

[15] Cf. Bostrom (2006).

[16] One could reverse the question and look instead at possible reasons for a superintelligent singleton *not* to develop some technological capabilities. These include: (a) The singleton foreseeing that it will have no use of some technological capability; (b) The development cost being too large relative to its anticipated utility. This would be the case if, for instance, the technology will never be suitable for achieving any of the singleton's ends, or if the singleton has a very high discount rate that strongly discourages investment; (c) The singleton having some final value that requires abstention from particular avenues of technology development; (d) If the singleton is not certain it will remain stable, it might prefer to refrain from developing technologies that could threaten its internal stability or that would make the consequences of dissolution worse (e.g., a world government may not wish to develop technologies that would facilitate rebellion, even if they had some good uses, nor develop technologies for the easy production of weapons of mass destruction which could wreak havoc if the world government were to dissolve); (e) Similarly, the singleton might have made some kind of binding strategic commitment not to develop some technology, a commitment that remains operative even if it would now be convenient to develop it. (Note, however, that some *current* reasons for technology-development would not apply to a singleton: e.g., reasons arising from unwanted arms races.)

On the basis of such observations it might be tempting to suppose that a superintelligence not facing a competitive social world would see no instrumental reason to accumulate resources beyond some modest level, for instance whatever computational resources needed to run its mind along with some virtual reality. Yet such a supposition would be entirely unwarranted. First, the value of resources depends on the uses to which they can be put, which in turn depends on the available technology. With mature technology, basic resources such as time, space, and matter, and other forms of free energy, could be processed to serve almost any goal. For instance, such basic resources could be converted into life. Increased computational resources could be used to run the superintelligence at a greater speed and for a longer duration, or to create additional physical or simulated (virtual) lives and civilizations. Extra physical resources could also be used to create backup systems or perimeter defenses, enhancing security. Such projects could easily consume far more than one planet's worth of resources.

Furthermore, the cost of acquiring additional extraterrestrial resources will decline radically as the technology matures. Once von Neumann probes can be built, a large portion of the observable universe (assuming it is uninhabited by intelligent life) could be gradually colonized—for the one-off cost of building and launching a single successful self-reproducing probe. This low cost of celestial resource acquisition would mean that such expansion could be worthwhile even if the value of the additional resources gained were somewhat marginal. For example, even if a superintelligence cared non-instrumentally only about what happens within some particular small volume of space, such as the space occupied by its original home planet, it would still have instrumental reasons to harvest the resources of the cosmos beyond. It could use those surplus resources to build computers to calculate more optimal ways of using resources within the small spatial region of primary concern. It could also use the extra resources to build ever-more robust defenses to safeguard the privileged real estate. Since the cost of acquiring additional resources would keep declining, this process of optimizing and increasing safeguards might well continue indefinitely even if it were subject to steeply declining returns.[17] [18]

---

[17] Suppose that an agent discounts resources obtained in the future at an exponential rate, and that because of the light speed limitation the agent can only increase its resource endowment at a polynomial rate. Would this mean that there will be some time after which the agent would not find it worthwhile to continue acquisitive expansion? No, because although the present value of the resources obtained at future times would asymptote to zero the further into the future we look, so would the present cost of obtaining them. The present cost of sending out one more von Neumann probe a 100 million years from now (possibly using some resource acquired some short time earlier) would be diminished by the same discount factor that would diminish the present value of the future resources the extra probe would acquire (modulo a constant factor).

[18] Even an agent that has an apparently very limited final goal, such as "to make 32 paperclips", could pursue unlimited resource acquisition if there were no relevant cost to the agent of doing so. For example, even after an expected-utility-maximizing agent had built 32 paperclips, it could use some extra resources to verify that it had indeed successfully built 32 paperclips meeting all the specifications (and, if necessary, to take corrective action). After it had done so, it could run another batch of tests to make doubly sure that no mistake had been made. And then it could run another test, and another. The benefits of subsequent tests would be subject to steeply diminishing returns; however, so long as there were no alternative action

Thus, there is an extremely wide range of possible final goals a superintelligent singleton could have that would generate the instrumental goal of unlimited resource acquisition. The likely manifestation of this would be the superintelligence's initiation of a colonization process that would expand in all directions using von Neumann probes. This would roughly result in a sphere of expanding infrastructure centered on the originating planet and growing in radius at some fraction of the speed of light; and the colonization of the universe would continue in this manner until the accelerating speed of cosmic expansion (a consequence of the positive cosmological constant) makes further material acquisition physically impossible as remoter regions drift permanently out of reach.[19] By contrast, agents lacking the technology required for inexpensive resource acquisition, or for the conversion of generic physical resources into useful infrastructure, may often find it not cost-effective to invest any present resources in increasing their material endowment. The same may hold for agents operating in competition with other agents of similar powers. For instance, if competing agents have already secured accessible cosmic resources, a late-starting agent may have no colonization opportunities. The convergent instrumental reasons for superintelligences uncertain of the non-existence of other powerful superintelligent agents are complicated by strategic considerations in ways that we do not currently fully comprehend but which may constitute important qualifications to the examples of convergent instrumental reasons we have looked at here.[20]

It should be emphasized that the existence of convergent instrumental reasons, even if they apply to and are recognized by a particular agent, does not imply that the agent's behavior is easily predictable. An agent might well think of ways of pursuing the relevant instrumental values that do not readily occur to us. This is especially true for a superintelligence, which could devise extremely clever but counterintuitive plans to realize its goals, possibly even exploiting as-yet undiscovered physical phenomena. What is predictable is that the convergent

---

with a higher expected utility, the agent would keep testing and re-testing (and keep acquiring more resources to enable these tests).

[19] While the volume reached by colonization probes at a given time might be roughly spherical and expanding with a rate proportional to the square of time elapsed since the first probe was launched ($\sim t^2$), the amount of resources contained within this volume will follow a less regular growth pattern, since the distribution of resources is inhomogeneous and varies over several scales. Initially, the growth rate might be $\sim t^2$ as the home planet is colonized; then the growth rate might become spiky as nearby planets and solar systems are colonized; then, as the roughly disc-shaped volume of the Milky Way gets filled out, the growth rate might even out, to be approximately proportional to t; then the growth rate might again become spiky as nearby galaxies are colonized; then the growth rate might again approximate $\sim t^2$ as expansion proceeds on a scale over which the distribution of galaxies is roughly homogeneous; then another period of spiky growth followed by smooth $\sim t^2$ growth as galactic superclusters are colonized; until ultimately the growth rate starts a final decline, eventually reaching zero as the expansion speed of the universe accelerates to such an extent as to make further colonization impossible.

[20] The simulation argument may be of particular importance in this context. A superintelligent agent may assign a significant probability to hypotheses according to which it lives in a computer simulation and its percept sequence is generated by another superintelligence, and this might various generate convergent instrumental reasons depending on the agent's guesses about what types of simulations it is most likely to be in. Cf. Bostrom (2003).

instrumental values would be pursued and used to realize the agent's final goals, not the specific actions that the agent would take to achieve this.

Conclusions

The orthogonality thesis suggests that we cannot blithely assume that a superintelligence will necessarily share any of the final values stereotypically associated with wisdom and intellectual development in humans—scientific curiosity, benevolent concern for others, spiritual enlightenment and contemplation, renunciation of material acquisitiveness, a taste for refined culture or for the simple pleasures in life, humility and selflessness, and so forth. It might be possible through deliberate effort to construct a superintelligence that values such things, or to build one that values human welfare, moral goodness, or any other complex purpose that its designers might want it to serve. But it is no less possible—and probably technically easier—to build a superintelligence that places final value on nothing but calculating the decimals of pi.

The instrumental convergence thesis suggests that we cannot blithely assume that a superintelligence with the final goal of calculating the decimals of pi (or making paperclips, or counting grains of sand) would limit its activities in such a way as to not materially infringe on human interests. An agent with such a final goal would have a convergent instrumental reason, in many situations, to acquire an unlimited amount of physical resources and, if possible, to eliminate potential threats to itself and its goal system.[21] It might be possible to set up a situation in which the optimal way for the agent to pursue these instrumental values (and thereby its final goals) is by promoting human welfare, acting morally, or serving some beneficial purpose as intended by its creators. However, if and when such an agent finds itself in a different situation, one in which it expects a greater number of decimals of pi to be calculated if it destroys the human species than if it continues to act cooperatively, its behavior would instantly take a sinister turn. This indicates a danger in relying on instrumental values as a guarantor of safe conduct in future artificial agents that are intended to become superintelligent and that might be able to leverage their superintelligence into extreme levels power and influence.[22]

References

Bostrom, N. (2003). Are You Living in a Computer Simulation? *Philosophical Quarterly*, 53(211), 243-255.

Bostrom. N. (2006). What is a Singleton? *Linguistic and Philosophical Investigations*, 5(2), 48-54.

[21] Human beings might constitute potential threats; they certainly constitute physical resources.

[22] For comments and discussion I am grateful to Stuart Armstrong, Grant Bartley, Owain Evans, Lisa Makros, Luke Muehlhauser, Toby Ord, Brian Rabkin, Rebecca Roache, Anders Sandberg, and three anonymous referees.

Bostrom, N. (2012). Information Hazards: A Typology of Potential Harms from Knowledge. *Review of Contemporary Philosophy*, 10, 44-79. [www.nickbostrom.com/information-hazards.pdf]

Chalmers, D. (2010): The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*, 17, 7-65.

Chislenko, A. (1997). Technology as Extension of Human Functional Architecture. *Extropy Online*. [project.cyberpunk.ru/idb/technology_as_extension.html]

de Blanc, P. (2011). Ontological Crises in Artificial Agent's Value Systems. *Manuscript*. The Singularity Institute for Artificial Intelligence. [arxiv.org/pdf/1105.3821v1.pdf]

Dewey, D. (2011). Learning What to Value. In Schmidhuber, J., Thorisson, K. R., Looks, M. (eds.). *Proceedings of the 4th Conference on Artificial General Intelligence*, AGI 2011 (pp. 309-314), Heidelberg: Springer.

Forgas, J. et al. (eds.) (2009). *The Psychology of Attitudes and Attitude Change*. London: Psychology Press.

Lewis, D. (1988). Desire as belief. *Mind*, 97(387), 323-332.

Omohundro, S. (2008a). The Basic AI Drives. In P. Wang, B. Goertzel, and S. Franklin (eds.). *Proceedings of the First AGI Conference*, 171, Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press.

Omohundro, S. (2008b). The Nature of Self-Improving Artificial Intelligence. *Manuscript*. [selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf]

Omohundro, S. (forthcoming 2012). Rationally-Shaped Artificial Intelligence. In Eden, A. et al. (eds.). *The Singularity Hypothesis: A Scientific and Philosophical Assessment* (Springer, forthcoming).

Parfit, D. (1984). *Reasons and Persons*. (pp. 123-4). Reprinted and corrected edition, 1987. Oxford: Oxford University Press.

Parfit, D. (2011). *On What Matters*. Oxford: Oxford University Press.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. (3rd ed.). New Jersey: Prentice Hall.

Sandberg, A. and Bostrom, N. (2008). *Whole Brain Emulation: A Roadmap*. Technical Report 2008–3. Oxford: Future of Humanity Institute, Oxford University.

[www.fhi.ox.ac.uk/Reports/2008-3.pdf]

Shulman, C. (2010). Omohundro's "Basic AI Drives" and Catastrophic Risks. *Manuscript*. [singinst.org/upload/ai-resource-drives.pdf]

Sinhababu, N. (2009). The Humean Theory of Motivation Reformulated and Defended. *Philosophical Review*, 118(4), 465-500.

Smith, M. (1987). The Humean Theory of Motivation. *Mind*, 46 (381): 36-61.

Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco: W. H. Freeman.

Yudkowsky, E. (2008). *Artificial Intelligence as a Positive and Negative Factor in Global Risk*. In Bostrom, N. and Cirkovic, M. (eds.). Global Catastrophic Risks. (pp. 308-345; quote from p. 310). Oxford: Oxford University Press.

Yudkowsky, E. (2011). Complex Value Systems Are Required to Realize Valuable Futures. In Schmidhuber, J., Thorisson, K. R., Looks, M. (eds.). *Proceedings of the 4th Conference on Artificial General Intelligence*, AGI 2011 (pp. 388-393). Heidelberg: Springer.