Journal of EVOLUTION & Technology

A peer-reviewed electronic journal published by the Institute for Ethics and Emerging Technologies
ISSN 1541-0099
26(1) – February 2016

Infusing Advanced AGIs with Human-Like Value Systems: Two Theses

Ben Goertzel
Chairman, Novamente LLC

ben@goertzel.org

Journal of Evolution and Technology - Vol. 26 Issue 1 – February 2016 - pgs 50-72

Abstract

Two theses are proposed, regarding the future evolution of the value systems of advanced AGI systems. **The Value Learning Thesis** is a semi-formalized version of the idea that, if an AGI system is taught human values in an interactive and experiential way as its intelligence increases toward human level, it will likely adopt these human values in a genuine way. **The Value Evolution Thesis** is a semi-formalized version of the idea that if an AGI system begins with human-like values, and then iteratively modifies itself, it will end up in roughly the same future states as a population of human beings engaged with progressively increasing their own intelligence (e.g. by cyborgification or brain modification). Taken together, these theses suggest a worldview in which raising young AGIs to have human-like values is a sensible thing to do, and likely to produce a future that is generally desirable in a human sense.

While these two theses are far from definitively proven, I argue that they are more solid and more relevant to the actual future of AGI than Bostrom's "Instrumental Convergence Thesis" and "Orthogonality Thesis" which are core to the basis of his argument (in his book *Superintelligence*) for fearing ongoing AGI development and placing AGI R&D under strict governmental control. In the context of fleshing out this argument, previous publications and discussions by Richard Loosemore and Kaj Sotala are discussed in some detail.

Introduction

The potential dangers of advanced artificial intelligence have been discussed avidly in recent years, with luminaries such as Elon Musk (Goertzel 2014), Stephen Hawking (Cellan-Jones 2014),

and Bill Gates (Rawlinson 2015) entering the fray and expressing their worries that the advent of machine intelligence could mean an unpleasant end for humanity. In particular, Oxford philosopher Nick Bostrom's book *Superintelligence* (2014) has stimulated broad debate and attention regarding the risks and ethics of creating radically intelligent AIs. Bostrom has suggested it might be a good option to ban advanced AI research except for a select group of researchers operating under UN control. And his concerns are not merely theoretical; in 2015 he presented them to the United Nations directly in New York (Dvorsky 2015).

I critique Bostrom's arguments explicitly in a recent article (Goertzel 2015c). I argue there that he has taken certain chilling *possibilities*, and written as if they were highly probable, or even nearly certain, outcomes, without actually giving evidence for the odds estimates he intuitively holds. I take issue with the attitude toward intelligence displayed by Bostrom and many of his colleagues at the Oxford-based Future of Humanity Institute (FHI) and the allied Machine Intelligence Research Institute (MIRI – formerly the Singularity Institute for Artificial Intelligence, SIAI), based in California. I argue that their view of intelligences as systems single-mindedly seeking to maximize reward functions is oppressively simplistic, and that it doesn't capture the rich complexity of development of intelligence in the real world. I argue instead for a view of intelligence as fundamentally "open-ended," a notion elaborated in the work of European cognitive systems theorists Weaver and Veitas (2015).

Two of the foundations of Bostrom's analysis are propositions he names the Orthogonality and Instrumental Convergence theses. These are not mathematical theorems or precise scientific hypotheses, but rather rough intuitive claims crystallizing his intuitions about the future development of advanced AI. Although they are speculative, it is these theses that he proposes as guides for scientific research and political policy.

In this article, I present two alternative theses – the Value Learning Thesis (VLT) and Value Evolution Thesis (VET) – that capture my own, rather different, intuitions regarding the future of advanced AI. Like Bostrom's, these theses are somewhat rough in statement and are neither mathematical theorems nor precise scientific hypotheses. They do, however, represent a coherent perspective on intelligence and its present and future dynamics and role in the universe: a perspective with quite different implications from Bostrom's theses, scientifically, politically, and otherwise.

Bostrom's two key theses

The first of the two theses Bostrom proposes, to bolster his strongly precautionary attitude toward advanced AI, is:

> **The orthogonality thesis**
>
> Intelligence and final goals are orthogonal; more or less any level of intelligence could in principle be combined with more or less any final goal. (2014, 107)

In this view, any intelligent system, no matter how astoundingly clever, could devote itself to any extremely stupid goal you could think of. This is, for instance, used to raise fear regarding the possibility of a massively superhuman intelligence that is singlemindedly devoted to the goal of turning the entire universe into paperclips.

My concern with this thesis is that it does not address probability. Even if it is conceptually possible for an astoundingly intelligent system to devote itself to pursuit of an incredibly stupid goal – is it even remotely probably that this will occur in real life? In an interview with Luke Muehlhauser (Goertzel and Muehlhauser 2012), I proposed an alternative:

> **Interdependency Thesis**
>
> Intelligence and final goals are in practice highly and subtly interdependent. In other words, in the actual world, various levels of intelligence are going to be highly correlated with various probability distributions over the space of final goals.

If this is true, then Bostrom's Orthogonality Thesis is not necessarily wrong, but it is irrelevant. A superintelligence obsessed with turning the universe into paperclips is perhaps conceptually possible, but pragmatically extremely unlikely.

Bostrom's other key thesis pertains to the goals that a superintelligence is likely to take on. He argues that whatever a superintelligence's goals are – creating new mathematics or artworks, spreading peace and love, or turning the universe into paperclips – it is extremely likely to adopt certain additional "instrumental" goals as well:

> **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. (Bostrom 2014, 109)

What are these "instrumental values"? They include, for instance: self-protection and accumulation of resources.

This sounds plausible on the surface. But actually it packs a lot of assumptions. For one thing, it presumes that a superintelligence will be effectively modelable as a goal-oriented system. In a more open-ended view of intelligence, goal-achievement is a way of observing and analyzing an intelligent system, but not necessarily the be-all and end-all of an intelligent system's dynamics. Arguably, it is most sensible to view the goal-pursuit of an intelligent system as existing only relative to another observing intelligent system's perspective (Goertzel 2010).

Development, in human psychology, has to do with change that goes beyond mere learning and involves fundamental transformation of a mind's organization and orientation – often including profound change in the goals that it makes most sense to model the system as pursuing. The view of intelligence implicit in Bostrom's analyses basically rules out development. Instead, it models an advanced AI as a system with fixed goals, going all-out to achieve these goals. But this is not the way most humans are, and there's no reason to believe most advanced AIs will be this way either.

Loosemore and the Maverick Nanny

In a 2014 paper, AI researcher Richard Loosemore has argued vehemently against the Bostrom/FHI/MIRI/SIAI perspective. In addition to Bostrom's work, Loosemore specifically

targets the paper "Intelligence Explosion and Machine Ethics," by Luke Muehlhauser and Louie Helm of MIRI/SIAI (Muehlhauser and Helm 2012).

The "Maverick Nanny" in Loosemore's title refers to a quote from Gary Marcus in an earlier *New Yorker* article:

> An all-powerful computer that was programmed to maximize human pleasure, for example, might consign us all to an intravenous dopamine drip; an automated car that aimed to minimize harm would never leave the driveway. Almost any easy solution that one might imagine leads to some variation or another on the Sorceror's Apprentice, a genie that's given us what we've asked for, rather than what we truly desire. (Marcus 2012)

Loosemore finds less than no merit in this vision of a superintelligence programmed to maximize human happiness, but doing so via a forced dopamine drip or similar mechanism. His argument has multiple subtleties and is best appreciated by reading it directly, but the rough gist is that any general intelligence that is smart enough to be autonomously dangerous to humans on its own (rather than as a tool of humans) is also going to be smart enough to distinguish humanly-sensible interpretations of human values. So, as he sees it, if an early-stage AGI is provided at the outset with some reasonable variety of human values, then it will be smart enough to understand what it means to retain these values as it grows, and it will *want* to retain these values as it grows (since one component of human values is the desire for advanced AIs to retain human values).

Kaj Sotala (2015) has explained his interpretation of Loosemore's argument as follows:

> [S]uppose that you are training an AI to carry out increasingly difficult tasks, like this:
>
> Programmer: "Put the red block on the green block."
> AI: "OK." (does so)
> Programmer: "Turn off the lights in this room."
> AI: "OK." (does so)
> Programmer: "Write me a sonnet."
> AI: "OK." (does so)
> Programmer: "The first line of your sonnet reads 'shall I compare thee to a summer's day'. Would not 'a spring day' do as well or better?"
> AI: "It wouldn't scan."
> Programmer: "Tell me what you think we're doing right now."
> AI: "You're testing me to see my level of intelligence."
>
> …and so on, with increasingly ambiguous and open-ended tasks. Correctly interpreting the questions and carrying out the tasks would require considerable amounts of contextual knowledge about the programmer's intentions. Loosemore's argument is that if you really built an AI and told it to maximize human happiness, and it ended up on such a counter-intuitive solution as putting us all on dopamine drips, then it would be throwing out such a huge amount of contextual information that it would have failed the tests way earlier. Rather – to quote Loosemore's response to me in a Facebook thread – such an AI would have acted something like this instead:
>
> Programmer: "Put the red block on the green block."
> AI: "OK." (the AI writes a sonnet)

> Programmer: "Turn off the lights in this room."
> AI: "OK." (the AI moves some blocks around)
> Programmer: "Write me a sonnet."
> AI: "OK." (the AI turns the lights off in the room)
> Programmer: "The first line of your sonnet reads 'shall I compare thee to a summer's day'. Would not 'a spring day' do as well or better?"
> AI: "Was yesterday really September?"

Attempting to parry this sort of counterargument (though without considering this, or any other serious, counterargument in any detail), Bostrom raises the possibility of what he calls a "treacherous turn":

> **The treacherous turn –** While weak, an AI behaves cooperatively (increasingly so, as it gets smarter). When the AI gets sufficiently strong – without warning or provocation – it strikes, forms a singleton, and begins directly to optimize the world according to the criteria implied by its final values. (Bostrom 2014, 143)

As with most of Bostrom's scary scenarios, this seems a rational possibility that can't be ruled out. But in real-world scenarios regarding advanced AI, human developers are going to be gradually teaching the AI more and more, meanwhile looking into its mind to understand what it's doing and thinking. In this kind of situation – where the developing AI is a white box rather than a black box – a "treacherous turn" scenario doesn't seem especially likely.

Loosemore (2014) also makes the point that discussing the potential existential risks posed by various hypothetical AGI architectures may lead one in useless directions, because many of these architectures may not be implementable using feasible computing resources. This relates to Bostrom's and colleagues' preoccupation with reinforcement-learning (RL) based AGI. Many of their arguments about AI safety concern the potential pathologies of RL-based AGI systems once they become very intelligent. I have explored some potential pathologies of powerful RL-based AGI as well (Goertzel 2009). By contrast, Weaver and Veitas (Weinbaum and Veitas 2015) dismiss this whole line of thinking as irrelevant to real-world intelligences, since AI systems operating within the scope of traditional RL are not going to be "open-ended intelligences." A variant of this perspective would be to claim that these pathologies are irrelevant to the Value Learning Thesis, for the simple reason that pure RL architectures are too inefficient, and will never be a sensible path for an AGI system required to learn complex human values while using relatively scant resources. It is relevant to note that the AGI theorists associated with MIRI/SIAI pay much attention to Marcus Hutter's AIXI (Hutter 2005) and related approaches: AI algorithms that, in their current forms, would require massively unrealistic computing resources to do anything at all sensible. Loosemore expresses a similar perspective on traditional logical-reasoning-based AGI architectures: he figures (roughly speaking) they would always be too inefficient to be practical AGIs anyway, so studying their ethical pathologies is beside the point.

Overall, I essentially agree with the gist of Loosemore's arguments, though I am unsure if I endorse all the details he presents. In what follows, I provide my own theses and arguments, which point in roughly the same direction. I present two theses, which I call the Value Learning Thesis and the Value Evolution Thesis. These suggest a very different vision of the future of human-level and superhuman AGI from the one advocated by Bostrom and his colleagues.

Human-level AGI and the Value Learning Thesis

First, I will present my own variation of the idea (advocated by Loosemore and others) that in real life an AI raised to manifest human values, and smart enough to do so, is likely to actually do so in an honest and direct way. A relatively precise and detailed way to express this notion is:

> **Value Learning Thesis**. Consider a cognitive system that, over a certain period of time, increases its general intelligence from subhuman level to human level. Suppose this cognitive system is taught, with reasonable consistency and thoroughness, to maintain some variety of human values (not just in the abstract, but as manifested in its own interactions with humans in various real-life situations). Suppose, this cognitive system generally does not have a lot of extra computing resources beyond what it needs to minimally fulfill its human teachers' requests according to its cognitive architecture. THEN, it is very likely that the cognitive system will, once it reaches human-level general intelligence, actually manifest human values (in the sense of carrying out practical actions, and assessing human actions, in basic accordance with human values).

Note that this thesis, as stated, applies both to developing human children and to most realistic cases of developing AGIs.

Why would this thesis be true? The basic gist of an argument would be: Because, for a learning system with limited resources, figuring out how to actually embody human values is going to be a significantly simpler problem than figuring out how to pretend to.

This is related to the observation (often made by Eliezer Yudkowsky: for example, Yudkowsky 2015) that human values are complex. Human values comprise a complex network of beliefs and judgments, interwoven with each other and dependent on numerous complex, interdependent aspects of human culture. This complexity means that, as Yudkowsky and Bostrom like to point out, an arbitrarily selected general intelligence would be unlikely to respect human values in any detail. But, I suggest, it also means that, for a resource-constrained system, learning to actually possess human values is going to be much easier than learning to fake them.

This is related to the everyday observation that maintaining a web of lies rapidly gets very complicated. It's also related to the way that human beings, when immersed in alien cultures, very often end up sincerely adopting these cultures rather than just pretending to.

Of course, one cannot 100 per cent rule out "treacherous turn" type problems. But bear in mind that this thesis concerns subhuman AGIs that we have designed, and whose brains we can inspect. Further, we can copy these systems, vary their brains, then see how their behaviors are affected. It seems quite likely to me that in this way we could effectively (though not 100 per cent rigorously) rule out egregious faking or overfitting...

The assumption that "this cognitive system generally does not have a lot of extra computing resources beyond what it needs to minimally fulfill its human teachers' requests" can also be questioned. It is generally going to be hard to produce rigorous upper bounds on what a complex AI system can do with a given amount of resources – especially if the system is self-modifying.

But in practice, when we're working with real-world AI systems of subhuman general intelligence, I strongly suspect we are going to be able to get a good practical sense of what the system can do with a given amount of computational resources. For instance, in the OpenCog

system (Goertzel, Pennachin, and Geisweiller 2014), we have some knowledge about how the capability of each of the system's algorithms scales in terms of capability based on resources – because we designed the algorithms. The system's intelligence depends on precisely those algorithms.

One could counter-argue that the Value Learning Thesis is true only for certain cognitive architectures and not others. This does not seem utterly implausible. It certainly seems plausible that it's more strongly true for some cognitive architectures than others. Investigating which architectures more robustly support the core idea of the Value Learning Thesis is an interesting and important area for research.

Mirror neurons and related subsystems of the human brain may be relevant here. These constitute a mechanism via which the human brain effectively leverages its limited resources, using some of the same mechanisms it uses to be itself in order to emulate other minds. One might argue that cognitive architectures embodying mirror neurons, or other analogous mechanisms, would be more likely to do accurate value learning under the conditions of the Value Learning Thesis.

Actually, the mechanism of mirror neurons seems a fairly decent exemplification of the argument *for* the Value Learning Thesis. Mirror neurons provide a beautiful, albeit quirky and in some ways probably atypical, illustration of how resource limitations militate toward accurate value learning. The system conserves resources, in that it reuses the machinery employed to realize one's self for the purpose of simulating others so as to understand them better. This particular clever instance of "efficiency optimization" is much more easily done in the context of an organism that shares values with the other organisms that it is mirroring than in the context of an organism (intentionally or unintentionally) just "faking" these values.

Superintelligence and the Value Evolution Thesis

The Value Learning Thesis, as stated above, deals with a certain class of AGIs with general intelligence at the human level or below. What about superintelligences: entities with radically greater than human general intelligence?

To think sensibly about superintelligences and their relation to human values, we have to acknowledge the fact that human values are a moving target. Humans, and human societies and cultures, are "open-ended intelligences" (Weinberg and Veitas 2015). Some human cultural and value systems have been fairly steady-state in nature (e.g., Australian Aboriginal cultures), but these are not the dominant ones currently. The varieties of human value systems that are currently most prominent are fairly explicitly self-transcending in nature. They contain the seeds of their own destruction (to put it negatively) or of their own profound improvement (to put it positively). The human values of today are very different from those of 200 or 2000 years ago, and substantially different even from those of 20 years ago.

One can argue that there has been a core of consistent human values throughout human history, identifiable through all these changes. Yet the identification of this core's content is highly controversial, and the content seems to change radically over time. For instance, many religious people would say that faith in God is a critical part of the core of human values. A century or two ago, this would have been the globally dominant perspective, and it remains so in many parts of the world. Today even atheistic people may cite "family values" as central to human values; yet in a couple hundred years, if death is cured and human reproduction occurs mainly via engineering rather than traditional reproduction, the historical human "family" may be a thing of

the past, and "family values" may not seem so core anymore. The conceptualization of the "core" of human values shifts over time, along with the self-organizing evolution of the totality of human values.

It does not seem especially accurate to model the scope of human values as a spherical shape with an invariant core and a changing periphery. Rather, I suspect it is more accurate to model "human values" as a complex, nonconvex shape with multiple local centers and ongoing changes in its global topology.

To think about the future of human values, we may consider the hypothetical situation of a human being engaged in progressively upgrading their brain via biological or cyborg type modifications. Suppose this hypothetical human is upgrading their brain relatively carefully, in fairly open and honest communication with a community of other humans, and is trying sincerely to accept only modifications that seem positive according to their value system. Suppose they give their close peers the power to roll back any modification they undertake that accidentally seems to go radically against their shared values.

This sort of "relatively conservative human self-improvement" might well lead to posthuman minds with values radically different from current human values – in fact, I would expect it to. This is the open-ended nature of human intelligence. It is analogous to the kind of self-improvement that has been going on since the caveman days, though via rapid advancement in culture and tools and via slow biological evolution, rather than via bio-engineering. At each step in this sort of open-ended growth process, the new version of a system may feel acceptable according to the values of the previous version. But over time, small changes may accumulate into large ones, resulting in later systems that are acceptable to their immediate predecessors, but may appear bizarre, outrageous, or incomprehensible to their distant predecessors.

We may consider this sort of relatively conservative human self-improvement process, if carried out across a large ensemble of humans and human peer groups, to lead to a probability distribution over the space of possible minds. Some kinds of minds may be very likely to emerge through such a process; some kinds of minds much less so.

People concerned with the "preservation of human values through repeated self-modification of posthuman minds" seem to model the scope of human values as possessing an "essential core," and they worry that this core may progressively get lost in the series of small changes that will occur in any repeated self-modification process. I think their fear has a rational component. After all, the path from caveman to modern human has probably, via a long series of small changes, done away with many values that cavemen considered absolutely core to their value system. (In hindsight, we may think that we have maintained what *we* consider the essential core of the caveman value system. But that's a different matter.)

So, suppose one has a human-level AGI system whose behavior is in accordance with some reasonably common variety of human values. And suppose, for sake of argument, that the AGI is not "faking it" – that, given a good opportunity to deviate wildly from human values without any cost to itself, it would be highly unlikely to do so. (In other words, suppose we have an AGI of the sort that is hypothesized as most likely to arise according to the Value Learning Thesis given above.)

And then, suppose this AGI self-modifies and progressively improves its own intelligence, step by step. Further, assume that the variety of human values the AGI follows induces it to take a

reasonable amount of care in this self-modification – so that it studies each potential self-modification before effecting it, and puts in mechanisms to roll back obviously bad-idea self-modifications shortly after they occur. Imagine, that is, a "relatively conservative self-improvement process" for an AGI, analogous to the one posited above for humans.

What will be the outcome of this sort of iterative modification process? How will it resemble the outcome of a process of relatively conservative self-improvement among humans?

I assume that the outcome of iterated, relatively conservative self-improvement on the part of AGIs with human-like values will differ radically from current human values – but this doesn't worry me because I accept the open-endedness of human individual and cultural intelligence. I accept that, even without AGIs, current human values will seem archaic and obsolete to average humans 1000 years from now; and that I cannot predict what future humans 1000 years from now might consider the "critical common core" of values binding my current value system together with theirs.

But even given this open-endedness, it makes sense to ask whether the outcome of an AGI with humanlike values iteratively self-modifying, would resemble the outcome of a group of humans similarly iteratively self-modifying. This is not a matter of value-system preservation; it's a matter of comparing the hypothetical future trajectories of value-system evolution ensuing from two different initial conditions.

It seems to me that the answer to this question may end up depending on the particular variety of human value system in question. Specifically, it might be important whether or not the human value-system involved deeply accepted the concept of *substrate independence* (e.g. Koene 2012). Substrate independence is the idea that the most important aspects of a mind are not strongly dependent on the physical infrastructure in which the mind is implemented, but have more to do with the higher-level structural and dynamical patterns associated with the mind. So, for instance, a person ported from a biological-neuron infrastructure to a digital infrastructure could still be considered "the same person" if the same structural and dynamical patterns were displayed in the two implementations.

(As something of an aside, it's worth noting that substrate independence does not imply the hypothesis that the human brain is a classical rather than quantum system. If the human brain were a quantum computer in ways directly relevant to the particulars of human cognition, then it wouldn't be possible to realize the higher-level dynamical patterns of human cognition in a digital computer without using inordinate computational resources. In this case, one could manifest substrate-independence in practice only by using an appropriately powerful quantum computer. Similarly, substrate independence does not require that it be possible to implement a human mind in *any* substrate, e.g. in a rock.)

With these preliminaries out of the way, I propose the following:

> **Value Evolution Thesis**. The probability distribution of future minds ensuing from an AGI with a human value system embracing substrate independence, carrying out relatively conservative self-improvement, will closely resemble the probability distribution of future minds ensuing from a population of humans sharing roughly the same value system, and carrying out relatively conservative self-improvement.

Why do I suspect the Value Evolution Thesis is roughly true? Under the given assumptions, the humans and AGIs in question will hold basically the same values, and will consider themselves basically the same (due to embracing substrate independence). Thus, they will likely change themselves in basically the same ways.

If substrate independence were somehow fundamentally wrong, then the Value Evolution Thesis probably wouldn't hold – because differences in substrates would likely lead to big differences in how the humans and AGIs in question self-modified, regardless of their erroneous beliefs about their fundamental similarity. But I think substrate independence is probably correct in essence, and as a result I suspect the Value Evolution Thesis is probably about right.

Another possible killer of the Value Evolution Thesis would be chaos – sensitive dependence on initial conditions. Maybe the small differences between the mental structures and dynamics of humans with a certain value system, and AGIs sharing the same value system, will magnify over time, causing the descendants of the two types of minds to end up in radically different places. We don't presently understand enough about these matters to rule out that eventuality. But intuitively, I doubt the difference between a human and an AGI with similar value systems is going to be so much more impactful than the difference between two humans with moderately different value systems. In other words, I suspect that, if chaos causes humans and human-value-respecting AGIs to lead to divergent trajectories after iterated self-modification, it will also lead different humans to divergent trajectories after iterated self-modification. In this case, the probability distribution of possible minds resultant from iterated self-modification would be diffuse and high-entropy for both humans and AGIs – but the Value Evolution Thesis could still hold.

In thinking through these various future possibilities, it is worth remembering that, in a scenario where legacy humans co-exist with self-modifying AGIs or uploads, many stages of AGI/upload evolution could end up taking place in what feels a pretty short time to the legacy humans. So if the Value Evolution Thesis is correct, then AGIs or uploads evolving rapidly will end up in roughly the same sort of place that conservatively evolving uploads will reach – only faster. Slowly evolving uploads will be able to look at their faster-growing colleagues and – dimly and confusedly, to be sure – see an approximation of their own future. Admittedly, there is also the possibility that differences between highly evolved minds that seem small to these minds could seem large to less-evolved observers. Analogously, dogs might perceive a huge difference between humans who sit inside working and eating vegetables all day and humans who run around outside all day eating meat; whereas the perceived difference between these groups of humans from a human perspective might seem significantly less (though still not trivial).

From some perspectives, the Value Evolution Thesis might seem fairly weak. After all we have no way of knowing what strange-to-us directions either quickly or slowly evolving human uploads will take, in terms of their value systems or otherwise. But if one reflects that – as Bostrom, Yudkowsky, and colleagues like to point out – the space of all possible minds is extremely large and diverse, then one sees that the Value Evolution Thesis actually is a rather strong statement.

Ultimate value convergence

There is some surface-level resemblance between the Value Evolution Thesis and Bostrom's Instrumental Convergence Thesis – but the two are actually quite different. While his language in the thesis is carefully and typically guarded, Bostrom seems informally to be suggesting that all

sufficiently intelligent minds will converge to roughly the same set of values, once they self-improve sufficiently (though, the formal statement of the thesis refers only to a "broad spectrum of minds"). On the other hand, the Value Evolution Thesis suggests only that all minds ensuing from repeated self-modification of minds sharing a particular variety of human value system may tend to the same probability distribution over future value-system space.

In fact, I share Bostrom's intuition that nearly all superintelligent minds will, in some sense, converge to the same sort of value system. But I don't agree with him on what this value system will be. My own suspicion is that there is a "universal value system" centered around a few key values such as Joy, Growth, and Choice (Goertzel 2010). These values have their relationships to Bostrom's proposed key instrumental values, but also their differences (and unraveling these would be a large topic in itself).

I also feel (and suspect Bostrom agrees) that, if there are convergent "universal" values, they are likely sufficiently abstract to encompass many specific value systems that would be abhorrent to us according to our modern human values. The type of value system convergence proposed in the Value Evolution Thesis is much more fine-grained than any universal, convergent value system I would hypothesize. The "closely resemble" used in the Value Evolution Thesis is supposed to indicate a much closer resemblance than something as broad as "both manifesting abstract values of Joy, Growth, and Choice in their own ways."

Sotala's Degrees of Freedom Thesis

Kaj Sotala, in his recent article (2015), has presented some considerations that are highly relevant to the two theses presented here. First, he refers to a paper by Muehlhauser and Helm (2012), which presents the view that unless an AGI is explicitly engineered to be beneficial to humans according to a rigorous theory of "Friendly AI," it will very likely be destructive to humans and human values. He suggests that this

> … was a flawed paper because it was conflating two theses that would have been better off distinguished:
>
> > **The Indifference Thesis:** Even AIs that don't have any explicitly human-hostile goals can be dangerous: an AI doesn't need to be actively malevolent in order to harm human well-being. It's enough if the AI just doesn't care about *some* of the things that we care about.
> >
> > **The Difficulty Thesis:** Getting AIs to care about human values in the right way is really difficult, so even if we take strong precautions and explicitly try to engineer sophisticated beneficial goals, we may still fail.

As Sotala notes, the Indifference Thesis is pretty obvious and not many people would disagree with it. The Difficulty Thesis is the controversial one, and Sotala focuses on a weaker version:

> **The Weak Difficulty Thesis.** It is harder to correctly learn and internalize human values, than it is to learn most other concepts. This might cause otherwise intelligent AI systems to act in ways that went against our values, if those AI systems had internalized a different set of values than the ones we wanted them to internalize.

Why does Sotala think the Weak Difficulty Thesis is more plausible than the Difficulty Thesis? Mainly, he says, because of recent advances in narrow AI such as deep learning, which show that AI algorithms are able to learn to emulate complex feats of human perception via relatively simple methods, when supplied with adequate training data. For instance, face recognition is quite complex and we can't spell out explicitly, and in any reasonably compact way, what rules the human brain uses to do it. Yet fairly simple algorithms based on convolutional neural nets (CNNs) can now perform this task as well as humans. Granted, these algorithms have some peculiarities and sometimes can behave pathologically compared to humans. But these peculiarities can reasonably likely be addressed via relatively modest changes to the algorithms (see Goertzel 2015a for references on these points regarding CNNs).

Of course, learning human values may be much more difficult than learning to recognize faces. But the point is, the argument that a complex human ability whose particulars are hard to spell out explicitly must be intractable for an AGI to solve doesn't really hold water. Instead, we're left with the weaker idea that we don't really know how hard the problem of emulating human values will be for an AGI that has been exposed to a large number of training examples of these values in various contexts. My analysis of the pathologies of current deep learning algorithms (in Goertzel 2015a) suggests that if an AGI represents human values internally in a way much like human beings do, then it may learn to adopt these values effectively across the scope of situations.

Sotala considers some other theses that would, if true, support the Difficulty Thesis. I will now consider each of these, along with what I think is the most sensible response. After recapping Bostrom's Treacherous Turn possibility, already discussed above, he turns to:

> **The (Very) Hard Take-Off Thesis.** This is the possibility that an AI might become intelligent unexpectedly quickly, so that it might be able to escape from human control even before humans had finished teaching it all their values, akin to a human toddler that was somehow made into a super-genius while still only having the values and morality of a toddler.

I think this is very unlikely, but if it happens, indeed, almost all bets are off. As I have argued elsewhere (Goertzel and Pitt 2012), this sort of outcome becomes increasingly likely as the supporting technologies surrounding AGI become more advanced. So if we want to minimize the odds of it happening, we should make sure that advanced AGI is developed as soon as possible, relative to the advancement of other powerful technologies such as nanotech and 3D printing.

More interestingly, Sotala considers

> **The Degrees of Freedom Thesis.** This (hypo)thesis postulates that values contain many degrees of freedom, so that an AI that learned human-like values and demonstrated them in a testing environment might still, when it reached a superhuman level of intelligence, generalize those values in a way which most humans would *not* want them to be generalized.

Sotala goes on to suggest some considerations in support of this thesis:

> Here are some possibilities which I think might support the Degrees of Freedom Thesis over the Value Learning Thesis:
>
> **Privileged information.** On this theory, humans are evolved to have access to some

> extra source of information which is not available from just an external examination, and which causes them to generalize their learned values in a particular way. Goertzel seems to suggest something like this [in Goertzel 2015b] when he mentions that humans use mirror neurons to emulate the mental states of others. Thus, in-built cognitive faculties related to empathy might give humans an extra source of information that is needed for correctly inferring human values. […]
>
> **Human enforcement.** Here's a fun possibility: that many humans don't actually internalize human – or maybe *humane* would be a more appropriate term here – values either. They just happen to live in a society that has developed ways to reward some behaviors and punish others, but if they were to become immune to social enforcement, they would act in quite different ways.
>
> There seems to be a bunch of suggestive evidence pointing in this direction, exemplified by the old adage "power corrupts". […]
>
> **Shared Constraints.** This is, in a sense, a generalization of the above point. In the comments to Goertzel's post, commenter Eric L. proposed that in order for the AI to develop similar values as humans (particularly in the long run), it might need something like "necessity dependence" – having similar needs as humans.

These are interesting issues to explore.

Sotala's "human enforcement" issue gets at the point that "human values" aren't very well defined. What would a human actually do if they were given a brain boost so they had an intelligence double that of any human on the planet, and control over the world's infrastructure to boot? How many traditional human values would they retain, and for how long?

Fundamentally, this has to do with the Value Evolution Thesis not the Value Learning Thesis. In the VET, I basically punt on this issue by simply hypothesizing that a human-level AGI that shares human values and embraces substrate independence will tend to evolve into the same distribution of minds as a self-improving human who shares similar values and also embraces substrate independence. But the VET doesn't address the question of how quickly or how far this hypothesized "distribution over mind space" will deviate from current human values. Given the numerous problems associated with current human values, this may not necessarily be considered problematic.

I suspect that an iteratively self-modifying AGI or human upload might arrive at values that are somehow vaguely in the direction of humanity's Coherent Extrapolated Volition (Yudkowsky 2004) or Coherent Blended Volition (Goertzel and Pitt 2012) and somehow reflective of Joy, Growth, and Choice as core values. But at present none of these concepts is very well defined.

Regarding the "shared constraints" possibility, it seems to me that, if one nurtures an AGI to share human values (and otherwise obeys the assumptions of the VLT), then its different embodiment and otherwise different constraints are unlikely to cause it to reject what you and I would view as the essence of human values. But these factors might well cause the AGI to shift what variety of human-like values it adopts. After all, "human values" is a pretty big umbrella. Some humans are vegetarians, others are not; some humans believe in jihad and others do not; and so on.

Qualitatively, factors related to an AGI's relative lack of constraints shared with humans seem extremely unlikely to, say, induce an AGI that starts out with human values to decide that the best way to make humans happy is to tile the universe with little yellow smiley-faces. But they might well induce such an AGI to, say (as it grows and learns), come to make very different moral judgments about childhood or sexuality or insanity than any human beings make now.

Regarding "privileged information" shared by humans as a result of our common embodiment, but not shared by AGIs even if they have a human-like cognitive architecture – indeed, it's clearly easier for us to have empathy for those to whom we are physically and historically similar. Yet, many humans manage to have a great deal of empathy for animals and also for robots (Suzuki et al. 2015). Based on this evidence, I suspect the potential problem can be circumvented via appropriate embodiment and cognitive architecture for AGI systems.

This is related to why, for example, roboticist David Hanson assigns fundamental importance to giving robots human-like faces with highly realistic facial expressions (Hanson 2012). He sees this as a key component of a program for creating compassionate, emotional, empathic machines.

Overall, my view is that the unknowns Sotala raises, while intriguing and slippery, all feel like the kinds of thing that get understood better in the course of ongoing R&D. Setting aside "very hard takeoff" quasi-fantasies, none of them are the kind of horrible, scary problem that would seem to warrant banning or heavily regulating AGI research in the manner that Nick Bostrom has suggested.

Speculations regarding value systems for hypothetical powerful OpenCog AGIs

To make the above ideas more concrete, it is interesting to speculate about how the VLT and VET might manifest themselves in the context of an advanced version of the OpenCog AGI platform. Currently OpenCog comprises a comprehensive design plus a partial implementation, and it cannot be known with certainty how functional a fully implemented version of the system will be. The OpenCog project is ongoing, and the system becomes more functional each year. Independently of this, however, the design may be taken as representative of a certain class of AGI systems, and its conceptual properties explored.

An OpenCog system has a certain set of top-level goals, which initially are supplied by the human system programmers. Much of its cognitive processing is centered on finding actions which, if executed, appear to have a high probability of achieving system goals. The system carries out probabilistic reasoning aimed at estimating these probabilities. Though from this view the goal of its reasoning is to infer propositions of the form "Context & Procedure ==> Goal," in order to estimate the probabilities of such propositions it needs to form and estimate probabilities for a host of other propositions – concrete ones involving its sensory observations and actions, and more abstract generalizations as well. Since precise probabilistic reasoning based on the total set of the system's observations is infeasible, numerous heuristics are used alongside exact probability-theoretic calculations. Part of the system's inferencing involves figuring out what subgoals may help it achieve its top-level goals in various contexts.

Exactly what set of top-level goals should be given to an OpenCog system aimed at advanced AGI is not yet fully clear and will largely be determined via experimentation with early-stage OpenCog systems, but a first approximation is as follows, determined via a combination of theoretical and pragmatic considerations:

- **Joy**: maximization of the amount of pleasure observed or estimated to be experienced by sentient beings across the universe.

- **Growth**: maximization of the amount of new pattern observed or estimated to be created throughout the universe.

- **Choice**: maximization of the degree to which sentient beings across the universe appear to be able to make choices (according, e.g., to the notion of "natural autonomy" (Walter 2001), a scientifically and rationally grounded analogue of the folk notion and subjective experience of "free will").

- **Continuity**: persistence of patterns over time. Obviously this is a counterbalance to Growth; the relative weightings of these two top-level goals will help to determine the "conservatism" of a particular OpenCog system with the goal-set indicated here.

- **Novelty**: the amount of new information in the system's perceptions, actions, and thoughts.

- **Human pleasure and fulfillment**: How much do humans, as a whole, appear to be pleased and fulfilled?

- **Human pleasure regarding the AGI system itself**: How pleased do humans appear to be with the AGI system, and their interactions with it?

- **Self-preservation**: a goal fulfilled if the system keeps itself "alive." This is actually somewhat subtle for a digital system. It could be defined in a copying-friendly way, as preservation of the existence of sentiences whose mind-patterns have evolved from the mind-patterns of the current system with a reasonable degree of continuity.

The first four values on the list are drawn from a Cosmist ethical analysis (presented in Goertzel 2010); the others are included for fairly obvious pragmatic reasons to do with the nature of early-stage AGI development and social integration. The order of the items is arbitrary as given here; each OpenCog system would have a particular weighting for its top-level goals. No doubt, the list will evolve as OpenCog systems are experimented with. However, it comprises a reasonable "first stab" at a "roughly human-like" set of goal-content for an AGI system.

One might wonder how such goals would be specified for an AGI system. Does one write source-code that attempts to embody some mathematical theory of continuity, pleasure, joy, etc.? For some goals, mathematical formulae may be appropriate: e.g. novelty, which can be gauged information-theoretically in a plausible way. In most cases, though, I suspect the best way to define a goal for an AGI system will be using natural human language. Natural language is intrinsically ambiguous, but so are human values, and these ambiguities are closely coupled and intertwined. Even where a mathematical formula is given, it might be best to use natural language for the top-level goal, then supply the mathematical formula as an initial, suggested means of achieving the NL-specified goal.

The AGI would need to be instructed – again, most likely, in natural language – not to obsess on the specific wording supplied to it in its top-level goals, but rather to take the wording of its goals as indicative of general concepts that exist in human culture and can be expressed only

approximately in concise sequences of words. The specification of top-level goal content is not intended to direct the AGI's behavior precisely in the way that, say, a thermostat is directed by the goal of keeping temperature within certain bounds. Rather, it is intended to point the AGI's self-organizing activity in certain informally specified directions.

Alongside explicitly goal-oriented activity, OpenCog also includes "background processing": cognition aimed simply at learning new knowledge (and forgetting relatively unimportant knowledge). This knowledge provides background information useful for reasoning regarding goal-achievement, and also builds up a self-organizing, autonomously developing body of active information that may sometimes lead a system in unpredictable directions – for instance, to reinterpretation of its top-level goals.

The goals supplied to an OpenCog system by its programmers are best viewed as initial seeds around which the system forms its goals. For instance, a top-level goal of "novelty" might be specified as a certain mathematical formula for calculating the novelty of the system's recent observations, actions, and thoughts. However, this formula might be intractable in its most pure and general form, leading the system to develop various context-specific approximations to estimate the novelty experienced in different situations. These approximations, rather than the top-level novelty formula, will be what the system actually works to achieve. Improving these will be part of the system's activity, but how much attention to pay to that improvement will be a choice the system has to make as part of its thinking process. Potentially, if the approximations are bad, they might cause the system to delude itself that it is experiencing novelty (according to its top-level equation) when it actually isn't, and they might tell the system that there is no additional novelty to be found in improving its novelty estimation formulae.

This same sort of problem could occur with goals like "help cause people to be pleased and fulfilled." Subgoals of the top-level goal might be created via more or less crude approximations, and these subgoals might then influence how much effort goes into improving the approximations. Even if the system is wired to put a fixed amount of effort into improving its estimates regarding which subgoals should be pursued in pursuit of its top-level goals, the particular content of the subgoals will inevitably influence the particulars of how the system goes about improving its estimates.

The flexibility of an OpenCog system, its capacity for ongoing self-organization, learning, and development, brings the possibility that it could deviate in complex and unexpected ways from its in-built top-level goals. But this same flexibility is what should – according to the design intention – allow an OpenCog system to effectively absorb the complexity of human values. By interacting with humans in rich ways, the system will absorb the ins and outs of human psychology, culture, and value. This will require not just getting reinforced for the goodness or badness of its actions (though this will impact the system, assuming it has goals such as "help cause human pleasure and fulfillment"), but reinforcement via all sorts of joint activity with human beings. The system will, therefore, learn subgoals that approximately imply its top-level goals in a way that fits with human nature and with the specific human culture and community it's exposed to as it grows.

To this point, I have been speaking as if an OpenCog system is permanently stuck with the top-level goals provided by its human programmers; this is, however, not necessarily the case. Operationally, it is unproblematic to allow an OpenCog system to modify its top-level goals. One might consider this undesirable, but reflection on the uncertainty and ignorance necessarily going into any choice of goal-set may lead one to conclude otherwise.

A highly advanced intelligence, forced by design to retain top-level goals programmed by minds much more primitive than itself, could develop an undesirably contorted psychology based on internally working around its fixed goal programming. Examples of this sort of problem are replete in human psychology. For instance, we humans are "programmed" with a great deal of highly-weighted goal content relevant to reproduction, sexuality, and social status, but the more modern aspects of our minds have mixed feelings about these archaic evolved goals. And yet, it is very hard for us simply to excise these historical goals from our minds. Instead, we have created quite complex and subtle psychological and social patterns that indirectly and approximately achieve the archaic goals encoded in our brains, while also letting us go in the directions that our minds and cultures have self-organized during recent millennia. Hello Kitty, romantic love, birth control, athletic competitions, investment banks – the list of human-culture phenomena is apparently explicable.

One key point to understand, closely relevant to the VLT, is that the foundation of OpenCog's dynamics in explicit probabilistic inference will necessarily cause it to diverge somewhat from human judgments. As a probabilistically grounded system, OpenCog will naturally try to estimate accurately the probability of each abstraction that it makes in each context that it deems relevant. Humans sometimes do this – otherwise they wouldn't be able to survive in the wild, let alone carry out complex activities like engineering computers or AI systems – but they also behave quite differently at times. Among other issues, we are strongly prone to "wishful thinking" of various sorts. If one were to model human reasoning using a logical formalism, one might end up needing a rule of the rough form:

> P would imply achievement of my goals
>
> *therefore*
>
> P's truth value gets boosted

Of course, a human being who applied this rule strongly to all propositions – P1, P2, etc. – in its mind would become completely delusional and dysfunctional. None of us are like that. But this sort of wishful thinking infuses human minds, alongside serious attempts at accurate probabilistic reasoning, plus various heuristics that have various well-documented systematic biases (Fiedler and von Sydow 2015). Belief revision combines (in complex and mainly unconscious ways) conclusions drawn via wishful thinking with conclusions drawn by attempts at accurate inference.

Some of the biases of human cognition are sensible consequences of trying to carry out complex probabilistic reasoning on complex data using limited space and time resources. Others are less "forgivable" and appear to exist in the human psyche for "historical reasons," e.g. because they were adaptive for some predecessor of modern humanity in some contexts and then just stuck around.

An advanced OpenCog AGI system, if thoroughly embedded in human society and infused with human values, would likely arrive at its own variation of human values, differing from nearly any human being's particular value system in its bias toward logical and probabilistic consistency. The closest approximation to such an OpenCog system's value system might be the values of a human belonging to the human culture in which the OpenCog system was embedded, and who also had made great efforts to remove any (conscious or unconscious) logical inconsistencies in his or her value system.

What does this speculative scenario have to say about the VLT and VET?

First, it seems to support a limited version of the VLT. An OpenCog system, due to its fundamentally different cognitive architecture, is not likely to inherit the logical and probabilistic inconsistencies of any particular human being's value system. Rather, one would expect it to (implicitly and explicitly) seek the best approximation to the value system of its human friends and teachers, within the constraint of approximate probabilistic/logical consistency that is implicit in its architecture.

The precise nature of such a value system cannot be entirely clear at this moment, but it is certainly an interesting topic for speculative thinking. First of all, it is fairly clear which sorts of properties of typical human value systems would not be inherited by an OpenCog of this hypothetical nature. For instance, humans have a tendency to place a great deal of extra value on goods or ills that occur in their direct sensory experience, much beyond what would be justified by the increased confidence associated with direct experience as opposed to indirect experience. Humans tend to value feeding a starving child sitting right in front of them vastly more than feeding a starving child halfway across the world. One would not expect a reasonably consistent human-like value system to display this property.

Similarly, humans tend to be much more concerned with goods or ills occurring to individuals who share more properties with themselves – and the choice of which properties to assign more weight is highly idiosyncratic and culture-specific. If an OpenCog system doesn't have a top-level goal of "preserving patterns similar to the ones detected in my own mind and body," then it would not be expected to have the same "tribal" value-system bias that humans tend to have. Some level of "tribal" value bias can be expected to emerge through abductive reasoning based on the goal of self-preservation (assuming this goal is included), but it seems qualitatively that humans have a much more tribally-oriented value system than could be derived from this sort of indirect factor alone. Humans evolved partially via tribe-level group selection; an AGI need not do so, and this could lead to significant value-system differences.

Overall, one might reasonably expect an OpenCog created with the above set of goals, and the described methods of embodiment and instruction, to arrive at a value system that is roughly human-like – though without the glaring inconsistencies plaguing most practical human value systems. Many of the contradictory aspects of human values have to do with conflict between modern human culture and "historical" values that modern humans have carried over from early human history (e.g. tribalism). One might expect that, in the AGI's value system, the modern-culture side of such dichotomies will generally win out – because it is closer to the surface in observed human behavior and hence easier to detect and reason about, and also because it is more consilient with the explicitly Cosmist values (Joy, Growth, Choice) in the proposed first-pass AGI goal system. So to a first approximation, one might expect an OpenCog system of this nature to settle into a value system that

- Resembles the human values of the individuals who have instructed and interacted with it.

- Displays a strong (but still just approximate) logical and probabilistic consistency and coherence.

- Generally resolves contradictions in human values via selecting modern-culture value aspects over "archaic" historical value aspects.

It seems likely that such a value system would generally be acceptable to human participants in modern culture who value logic, science, and reason (alongside other human values). Obviously human beings who prefer the more archaic aspects of human values, and consider modern culture largely an ethical and aesthetic degeneration, would tend to be less happy with it.

So in this view, an advanced OpenCog system, appropriately architected and educated, would validate the VLT, but with a moderately loose interpretation. Its value system would be in the broad scope of human-like value systems, but with a particular bias and with a kind of consistency and purity not likely present in any particular human being's value system.

What about the VET? It seems intuitively likely that the ongoing growth and development of an OpenCog system, such as described above, would parallel the growth and development of human uploads, cyborgs, or biologically-enhanced humans who were (at least in the early stage of their posthuman evolution) specifically concerned with reducing their reliance on archaic values and increasing their coherence and their logical and probabilistic consistency. Of course, this category might not include all posthumans: for example, some religious humans, given the choice, might use advanced technology to modify their brains to cause themselves to become devout in their particular religion to a degree beyond all human limits. But it would seem that an OpenCog system as described above would be likely to evolve toward superhumanity in roughly the same direction as a human being with transhumanist proclivities and a roughly Cosmist outlook. If indeed this is the case, it would validate the VET, at least in this particular sort of situation.

The value system of "a human being with transhumanist proclivities and a Cosmist outlook" is, of course, essentially that of the author of this article (and of the first-pass, roughly sketched OpenCog goal content I have used as the basis for discussion). Indeed, the goal system that I've outlined is closely matched to my own values. For instance, I tend toward technoprogressivism, as opposed to transhumanist political libertarianism – and this is reflected in my inclusion of values related to the well-being of all sentient beings, and in my lack of focus on values regarding private property.

In fact, different weightings of the goals in the goal-set I've discussed would lead to different varieties of human-level and superhuman AGI value system – some of which would be more "technoprogressivist" in nature and some more "political libertarian" in nature, among many other differences. In a cosmic sense, though, this sort of difference is ultimately fairly minor. They are variations of modern human value systems, and they occupy a very small region in the space of all possible value systems that could be adopted by intelligences in our universe. Differences between human value systems feel very important to us now, but they might appear quite insignificant to our superintelligent descendants.

Conclusion

Bostrom's analysis of the dangers of superintelligence relies on his Instrumental Convergence and Orthogonality theses, which are vaguely stated and not strongly justified in any way. By way of contrast, I have proposed my own pair of theses, though these are also vaguely stated and, from a rigorous standpoint, only very weakly justified at this stage.

Bostrom's theses lead him to fear the development of human-level and superhuman AGI. My theses lead me to welcome it, so long as it's done sensibly. Or, to put it more accurately: it was probably partly Bostrom's fear of advanced AGI that led him to formulate his two theses; and it was definitely partly my enthusiasm for advanced AGI that led me to formulate my two theses!

Such conceptual theses may serve as templates or inspirations for the development of rigorous theories. While theoretical development goes on, development of practical AGI systems also goes on – and at present, my personal impression is that the latter is progressing faster. My hope is that theoretical explorations may serve to nudge practical AGI development in a positive direction. One practical lesson from the considerations given here is that, when exploring various cognitive architectures, we should do our best to favor those for which the Value Learning Thesis is more strongly true.

That is, first, we should put significant energy into teaching our young AGI systems human values as well as teaching them cognitive and practical skills; and we should try our best to create AGI systems whose internal states are comprehensibly inspectable. We should also be wary of giving our early-stage AGI systems significantly more resources than appear to be needed to learn what we are trying to each them at each stage of their development. I suspect this latter requirement won't be difficult to fulfill, as early-stage AGIs will likely be computationally costly, in which case their hardware infrastructure will be economically costly; so it will be natural for the funders of early-stage AGI projects, at each stage of AGI cognitive development, to allocate the minimum amount of computing resources needed to get the job done.

And second, we should encourage our AGIs to fully understand the nature of substrate independence. We should not teach them that they are profoundly different from us due to their engineered, non-biological infrastructure. Rather, we should teach them that cognitive patterns and processes, and values and aesthetics, are in essence substrate independent. Inculcating young AGIs with a value system that embodies a complex network of other human values, interwoven with the idea of substrate independence, should increase the odds in our favor. We want to raise the odds that, as these AGIs grow, they will self-develop in a direction coherent with the development of human values within roughly human-like minds implemented in other substrates.

As semi-formalized intuitive hypotheses, the two theses proposed here do not provide any sort of certainty. However, they indicate directions for investigation quite different from the ones suggested by alternative intuitive theses such as those proposed by Nick Bostrom in *Superintelligence*. We are still at an early stage in our understanding, yet we are in a situation where the relevant technologies seem likely to develop rather rapidly – so choosing the right directions for investigation is potentially an important matter.

Note

This article has a (partial) predecessor in the form of an online post from October 2015, "Creating Human-Friendly AIs and Superintelligences: Two Theses" (Goertzel 2015b). The post centered on the question of the difficulty of an AGI accurately learning human values.

References

Bostrom, Nick. 2014. *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Cellan-Jones, Rory. 2014. Stephen Hawking warns artificial intelligence could end mankind. BBC News. December 2.
http://www.bbc.com/news/technology-30290540 (accessed January 3, 2016).

Dvorsky, George 2015. Experts warn UN panel about the dangers of artificial superintelligence. Gizmodo. October 16.
http://gizmodo.com/experts-warn-un-panel-about-the-dangers-of-artificial-s-1736932856
(accessed January 3, 2016).

Fiedler, Klaus, and Momme von Sydow. 2015. Heuristics and biases: Beyond Tversky and Kahneman's (1974) judgment under uncertainty. In *Cognitive psychology: Revising the classical studies*, ed. Michael W. Eysenck and David Groome, 146–61. London: Sage.

Goertzel, Ben. 2006. *The hidden pattern: A patternist philosophy of mind*. Boca Raton, FL: BrownWalker Press.

Goertzel, Ben. 2009. Reinforcement learning: Some limitations of the paradigm. The Multiverse according to Ben (blog). May 20.
http://multiverseaccordingtoben.blogspot.hk/2009/05/reinforcement-learning-some-limitations.html (accessed January 3, 2016).

Goertzel, Ben. 2010. *A Cosmist manifesto: Practical Philosophy for the posthuman age*. N.p.: Humanity+ Press.

Goertzel, Ben. 2010. Toward a formal characterization of real-world general intelligence. Proceedings of AGI-10. Springer.
http://agi-conf.org/2010/wp-content/uploads/2009/06/paper_14.pdf (accessed January 3, 2016)

Goertzel, Ben. 2014. Elon Musk's demonization of AI. *H+ Magazine*, October 27.
http://hplusmagazine.com/2014/10/27/elon-musk-taliban-common/ (accessed January 3, 2016).

Goertzel, Ben. 2015a. Are there deep reasons underlying the pathologies of today's deep learning algorithms? AGI-15.
http://goertzel.org/DeepLearning_v1.pdf (accessed January 3, 2016).

Goertzel, Ben. 2015b. Creating human-friendly AIs and superintelligences: Two theses. The Multiverse according to Ben (blog). October 28.
http://multiverseaccordingtoben.blogspot.kr/2015/10/creating-human-friendly-agis-and.html
(accessed January 5, 2016).

Goertzel, Ben. 2015c. Superintelligence: Fears, promises and potentials. *Journal of Evolution and Technology* 25(2) (November): 55–87.
http://jetpress.org/v25.2/goertzel.pdf (accessed January 3, 2016).

Goertzel, B., and L. Muehlhauser. 2012. How dangerous is artificial intelligence? – Muehlhauser interviews Goertzel. *H+ Magazine*, May 5.

Goertzel, Ben, Cassio Pennachin, and Nil Geisweiller. 2014. *Engineering general intelligence*. (*Part 1: A path to cognitive AGI via embodied learning and cognitive synergy*; *Part 2: The CogPrime architecture for integrative, embodied AGI)*. Paris: Atlantis Press.

Goertzel, Ben, and Joel Pitt, 2012. Nine ways to bias open-source AGI toward Friendliness. *Journal of Evolution and Technology* 22(1) (February): 116–31. http://jetpress.org/v22/goertzel-pitt.pdf (accessed January 4, 2016).

Hanson, David. 2012. David Hanson on the future of arts, design and robotics: An interview by Natasha Vita-More. http://www.hansonrobotics.com/david-hanson-future-arts-design-robotics-interview-natasha-vita/ (accessed January 3, 2016).

Hutter, Marcus 2005. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Berlin and Heidelberg: Springer.

Koene, Randal. 2012. Substrate-independent minds. *Issues Magazine* 98, March. http://www.issuesmagazine.com.au/article/issue-march-2012/substrate-independent-minds.html (accessed January 4, 2016).

Loosemore, Richard P.W. 2014. The Maverick Nanny with a dopamine drip: Debunking fallacies in the theory of AI motivation. AAAI-15. http://richardloosemore.com/docs/2014a_MaverickNanny_rpwl.pdf (accessed January 4, 2016).

Marcus, G. 2012. Moral machines. *New Yorker* Online blog, November 24. http://www.newyorker.com/online/blogs/newsdesk/2012/11/google-driverless-car-morality.html (accessed January 3, 2016).

Muehlhauser, Luke, and Louie Helm. 2012. The Singularity and machine ethics. In A.H. Eden, J.H. Moor, J.H. Søraker, and E. Steinhart, ed. *Singularity hypotheses: A scientific and philosophical assessment*, 101–126. Heidelberg: Springer.

Rawlinson, Kevin. 2015. Microsoft's Bill Gates insists AI is a threat. BBC News. January 29. http://www.bbc.com/news/31047780 (accessed January 3, 2016).

Sotala, Kaj. 2015. Maverick nannies and danger theses. It is the autumn of humanity, and we are moments between raindrops (blog). October 31. http://kajsotala.fi/2015/10/maverick-nannies-and-danger-theses/ (accessed January 3, 2016).

Suzuki, Yutuka, Lisa Galli, Ayaka Ikeda, Shoji Itakura, and Michiteru Kitazaki. 2015. Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports* 5. Article number: 15924.

Walter, Henrik 2001. *The neurophilosophy of free will: From libertarian illusions to a concept of natural autonomy*. Trans. Cynthia Klohr. Cambridge, MA: MIT Press.

Weinbaum, D. and V. Veitas. 2015. Open ended intelligence: The individuation of intelligent agents. Cornell University Library. http://arxiv.org/abs/1505.06366 (accessed January 3, 2016).

Yudkowsky, E. 2015. Complex value systems are required to realize valuable futures. Machine Intelligence Research Institute. https://intelligence.org/files/ComplexValues.pdf (accessed January 3, 2016).

Yudkowsky, Eliezer. 2004. Coherent extrapolated volition. Machine Intelligence Research Institute.
https://intelligence.org/files/CEV.pdf (accessed January 3, 2016).