Journal of Evolution & Technology

A peer-reviewed electronic journal published by the Institute for Ethics and Emerging Technologies

ISSN 1541-0099

25(2) – December 2015

Superintelligence: Fears, Promises and Potentials

Reflections on Bostrom's *Superintelligence*, Yudkowsky's *From AI to Zombies,* and Weaver and Veitas's "Open-Ended Intelligence"

Ben Goertzel Ph.D.
Chairman, Novamente LLC

ben@goertzel.org

Journal of Evolution and Technology – Vol. 24 Issue 2 – November 2015 – pgs 55-87

Abstract

Oxford philosopher Nick Bostrom, in his recent and celebrated book *Superintelligence*, argues that advanced AI poses a potentially major existential risk to humanity, and that advanced AI development should be heavily regulated and perhaps even restricted to a small set of government-approved researchers. Bostrom's ideas and arguments are reviewed and explored in detail, and compared with the thinking of three other current thinkers on the nature and implications of AI: Eliezer Yudkowsky of the Machine Intelligence Research Institute (formerly Singularity Institute for AI), and David Weinbaum (Weaver) and Viktoras Veitas of the Global Brain Institute.

Relevant portions of Yudkowsky's book *Rationality: From AI to Zombies* are briefly reviewed, and it is found that nearly all the core ideas of Bostrom's work appeared previously or concurrently in Yudkowsky's thinking. However, Yudkowsky often presents these shared ideas in a more plain-spoken and extreme form, making clearer the essence of what is being claimed. For instance, the elitist strain of thinking that one sees in the background in Bostrom is plainly and openly articulated in Yudkowsky, with many of the same practical conclusions (e.g. that it may well be best if advanced AI is developed in secret by a small elite group).

Bostrom and Yudkowsky view intelligent systems through the lens of reinforcement learning – they view them as "reward-maximizers" and worry about what happens when a very powerful and intelligent reward-maximizer is paired with a goal system that gives rewards for achieving foolish goals like tiling the universe with paperclips. Weinbaum and Veitas's recent paper "Open-Ended Intelligence" presents a starkly alternative perspective on intelligence, viewing it as centered not on reward maximization, but rather on complex self-organization and self-transcending

development that occurs in close coupling with a complex environment that is also ongoingly self-organizing, in only partially knowable ways.

It is concluded that Bostrom and Yudkowsky's arguments for existential risk have some logical foundation, but are often presented in an exaggerated way. For instance, formal arguments whose implication is that the "worst case scenarios" for advanced AI development are extremely dire, are often informally discussed as if they demonstrated the likelihood, rather than just the possibility, of highly negative outcomes. And potential dangers of reward-maximizing AI are taken as problems with AI in general, rather than just as problems of the reward-maximization paradigm as an approach to building superintelligence. If one views past, current, and future intelligence as "open-ended," in the vernacular of Weaver and Veitas, the potential dangers no longer appear to loom so large, and one sees a future that is wide-open, complex and uncertain, just as it has always been.

1. Introduction

These times of obvious exponential technological acceleration are especially exciting for those of us who work on building thinking machines – what I call "Artificial General Intelligence" and others have sometimes called "Strong AI." In fact, the present scene confronts us AGI researchers with an intriguing situation. On the one hand, we are in a world newly filled with widespread optimism about the viability of achieving advanced AGI. On the other hand, we concurrently confront a minor epidemic of pessimism – at least in the media, and among a handful of high-profile commentators – about the desirability of achieving this goal.

To put it differently: In the last couple years, all of a sudden, when I tell someone I'm working on building AGIs with general capability at the human level and beyond, the first reaction is no longer inevitably "You're nuts, that's not possible now – maybe in ten thousand years," but is just as likely to be "Wait, didn't Google already do that last year?" And the second reaction is fairly likely to be, "But isn't Stephen Hawking worried that, if anyone does build a truly smart machine, it's going to kill everybody?"

The current state of scientific and technological development is still pretty far from human-level AGI. But we all know progress in science isn't generally linear, and recent practical successes of "narrow," application-specific AI programs seem to have – more so than prior AI successes – dramatically increased the world's general impression that we could possibly be nearing a period of rapid increase in the general intelligence of our AI systems.

And it is this apparent progress that has led science and technology luminaries like Elon Musk (Kumpara 2014), Stephen Hawking (Cellan-Jones 2014) and Bill Gates (Holley 2015) to publicly raise an alarm regarding the potential that, one day not necessarily that far off, superhuman AIs might emanate from some research lab and literally annihilate the human race. As Peter Norvig of Google has noted (personal communication), quite suddenly we've gone from outrage at the perceived failure of AI research, to outrage at the perceived possible success of AGI research!

One of the most influential recent voices to express worry about the possible consequences of AGI is that of Oxford philosopher Nick Bostrom, whose recent book *Superintelligence: Paths, Dangers, Strategies* (Bostrom 2014) has received plaudits far and wide. In fact, it seems that Bostrom's book may have been part of what raised the hackles of Gates, Musk, and Hawking regarding the potential near-to-medium term risks of advanced AI R&D.

Most of the themes of *Superintelligence* were already familiar to me before opening the book, due to having read so many of Nick Bostrom's papers in the past – along with a few face-to-face conversations with Nick, e.g. when we co-organized the AGI-Impacts conference at Oxford in 2012. However, none of Bostrom's previous writings have assembled his ideas on these themes nearly so systematically and cohesively. Clearly the book was the product of a long and careful effort.

In this moderately lengthy review-and-discussion article I summarize my reaction to Bostrom's book and ideas. A truly thorough analysis and criticism of Bostrom's themes would require a response longer than the book itself; here I restrict myself to what I see as the main points of *Superintelligence*, those pertaining directly to the perceived "existential risk" of advanced AI. This is the aspect of Bostrom's book that has attracted the most attention in the media and the intellectual sphere.

2. Scratch the surface of Bostrom, find Yudkowsky

My first reaction, as I paged with curiosity through Bostrom's gloomy, erudite tome, was to be struck by the remarkable similarity between the arguments, and the general attitude presented, and the perspective on the future of AI put forth by Eliezer Yudkowsky over the last 15 years or so. Yudkowsky has not (yet) written any single document as thorough and carefully-wrought as *Superintelligence*. But in mailing list and blog posts and online essays over the years, he has articulated a perspective very close to Bostrom's, and in many ways possessing more depth and nuance.

The situation seems broadly similar to the argument for the plausibility of a Technological Singularity that celebrated philosopher David Chalmers put forth in his article "The Singularity: A Philosophical Analysis", published in the *Journal of Consciousness Studies* in 2010 (Chalmers 2010). The core arguments presented by Chalmers were very familiar to knowledgeable readers from prior writings by Vernor Vinge, Ray Kurzweil and others. But these other authors had presented their arguments more casually and informally. Chalmers gave the argument for a Singularity a full-on analytic-philosophy treatment, which yielded some value. For instance, he made very clear that the vagueness of the concept of "intelligence" is no objection to the Vingean/Kurzweilian arguments for a Singularity. It's OK if "intelligence" is a confusingly weighted and interrelated grab-bag of different capabilities, so long as some of the key capabilities among this set have the right properties regarding being ongoingly pumpable via improvements in technology.

Similarly to Chalmers' excellent essay, there is not really much that's new conceptually in Bostrom's book. But what others (especially, but not only, Yudkowsky) previously put forward more informally or erratically, regarding the potential for existential risk from superintelligence, Bostrom has now articulated with the detail and intellectual force of a top-notch analytic philosopher. It helps that he is a very good writer, mixing academic rigor with occasional color and humor in a classic European style.

Less of an academic than Bostrom (he's never been to college, let alone acquired a Ph.D.) and more of a maverick, Yudkowsky has accumulated a substantial community of enthusiasts to help him develop and extend his ideas. Yudkowsky's Machine Intelligence Research Institute (previously Singularity Institute for AI) is edgier and less "establishment" than Bostrom's Oxford-based Future of Humanity Institute, but the two groups have strongly overlapping attitudes and mandates. Recently a large collection of Yudkowsky's blog posts have been gathered in a book titled *Rationality: From AI to Zombies* (Yudkowsky 2015). This book covers a variety of topics besides superintelligence, yet the majority of Yudkowsky's points tie into the superintelligence theme via one moderately direct route or another. For instance, Yudkowsky's long-standing preoccupation with "the art of rationality" was originally motivated substantially by a desire to make himself and others rational enough to think effectively about the path to friendly superintelligence. While his scope of interests is fairly wide-ranging (though with some specific, and in some cases telling, limits that will be noted below), the creation of human-friendly

superintelligence (or in his coinage, "Friendly AI") has been Yudkowsky's focus for the whole of his career and adult life.

Compared to Yudkowsky, Bostrom's tone is far more sober and academic. Yudkowsky – to summarize a bit too glibly – generally writes for young enthusiasts who want to feel like they're members of an exclusive club of super-brights who understand what's really going on in the world. Bostrom presents closely related ideas in a fashion tailored for the sober, buttoned-down grown-up world, spelling out complex arguments regarding themes that would normally reek of SF in such an impressively clear, reasoned and compelling way that the reader can't help but take them seriously.

In the remainder of this article, I will give some of my main reactions to Bostrom's and Yudkowsky's books; and I will also seek to present their ideas in the context of a broader perspective on superintelligence, and its promises and risks.

While my comments on Bostrom's and Yudkowsky's writings will sometimes be critical, I want to stress that I consider both Bostrom and Yudkowsky valuable commentators on the future of AI and humanity. More than that, I often find their perspectives delightful to encounter, even when I disagree with them! The future is hard to understand, and having a wide variety of smart, creative people thinking about it from various perspective, seems almost sure to be a Good Thing at the present stage.

3. Some important things Bostrom and Yudkowsky get right

On some key points, I feel Bostrom and Yudkowsky get it just right, whereas the mainstream of the AI research community and the tech media doesn't quite. They fully get the dynamics of exponential advance, and the ever present potential for sudden dramatic breakthroughs in cutting-edge science.

For instance, Microsoft founder Paul Allen argues that "the Singularity is far" (Allen 2011), and even in today's optimistic climate, I would bet that a substantial plurality or even a majority of AI professors at leading universities agree with him. Arguments with such folks tend to end up writhing around in a morass of confusing details, with nobody convinced to give up their initial position. Given the current state of AI technology, we really can't know exactly when "true AGI" or "human level AI" is going to emerge – and we also can't know that much about what the first true AGI systems are going to be like, so we can't say anything especially confident about the risks these systems may actually pose.

On the other hand, Singularity pessimists like Allen do seem to miss some of the points Bostrom so carefully outlines. Progress in science is often not linear. Often progress proceeds step by step for a while – and then some breakthrough happens, disrupting the state of the art and reorienting much of the field's efforts toward exploring and leveraging the breakthrough. Kurzweil likes to point out that exponential progress curves are made of cascades of "S" curves, a new one starting after the previous one.

From Bostrom's point of view, the Singularity optimists don't need to be fully right in order for superintelligence to be a major ethical concern. As he sees it, if human-level AGI is created it's very likely to lead fairly soon after to superintelligence; and if superintelligence is created without extremely particular conditions being met, then it's very likely to lead to terrible outcomes like the extermination of all humans. Therefore if there's even a 1 per cent chance that Singularity optimists are correct about the near advent of human-level AGI, there's a major risk worthy paying careful attention to.

Bostrom cites a number of surveys of AI experts regarding the likely time-horizon of the advent of advanced AGI, including a survey my father (Ted Goertzel), Seth Baum, and I did at the AGI-09 conference in Washington DC (Baum et al. 2010). Overall a majority of these experts expect human-level AGI this century, with a mean expectation around the middle of the century. My own predictions are

more on the optimistic side (one to two decades rather than three to four) but in the scheme of things this is a pretty minor difference.

Bostrom notes that many of the experts surveyed expect a significant gap between the achievement of human-level AGI, and the emergence of radically superhuman AGI. He expects the gap to be smaller – more likely years than decades. Here also I tend to agree with him. Ray Kurzweil predicts the advance of human-level AI in 2029, and a general technological Singularity in 2045. On this I side with Nick Bostrom – I tend to think the gap between a human-level thinking machine and a world-transforming Singularity type event will be significantly less than 16 years.

Another point Bostrom hits home very well is the qualitative difference one can expect between the eventual superhuman AGIs we are likely to see emerge, and very smart human beings. As he puts it, using an example that also occurs in Yudkowsky's book and that I've seen in some of Yudkowsky's conference presentations over the years: It's not like the difference between Einstein and the village idiot, but more like the difference between the ultimately-fairly-similar minds of Einstein and the village idiot, and a mouse or a bacterium. Yes.

I tend to agree with Hugo de Garis (2005; Goertzel 2011) that anyone who thinks the end-game of AGI is going to be a mere modest improvement on human intelligence doesn't get the potential for intelligence inherent in the universe's available mass-energy.[1] The variety of intelligence that happens to have evolved in human brains appears to be nowhere near the maximally intelligent possible organization of mass-energy that the known laws of physics allow (not to mention physics as it will be understood by superintelligence…).

4. Staring into the Singularity

Yudkowsky understood the disruptive, frightening, exciting majesty of the Singularity pretty clearly way back in 1996 when he wrote the first version of his classic online essay "Staring into the Singularity" (parts of which he now repudiates; Yudkowsky 2001). "Staring" starts off with a bang:

> The short version:
>
> *If computing speeds double every two years,*
> *what happens when computer-based AIs are doing the research?*
>
> Computing speed doubles every two years.
> Computing speed doubles every two years of work.
> Computing speed doubles every two *subjective* years of work.
>
> Two years after Artificial Intelligences reach human equivalence, their speed doubles. *One* year later, their speed doubles again.
>
> Six months – three months – 1.5 months ... Singularity.
>
> Plug in the numbers for current computing speeds, the current doubling time, and an estimate for the raw processing power of the human brain, and the numbers match in: 2021.

[1] Though I should perhaps add that I don't agree with de Garis's various political prognostications.

But personally, I'd like to do it sooner.

1: The End of History

It began three and a half billion years ago in a pool of muck, when a molecule made a copy of itself and so became the ultimate ancestor of all earthly life.

It began four million years ago, when brain volumes began climbing rapidly in the hominid line.

Fifty thousand years ago with the rise of *Homo sapiens sapiens.*
Ten thousand years ago with the invention of civilization.
Five hundred years ago with the invention of the printing press.
Fifty years ago with the invention of the computer.

In less than thirty years, it will end.

At some point in the near future, someone will come up with a method of increasing the maximum intelligence on the planet – either coding a true Artificial Intelligence or enhancing human intelligence. An enhanced human would be better at thinking up ways of enhancing humans; would have an "increased capacity for invention". What would this increased ability be directed at? Creating the next generation of enhanced humans.

And what would those doubly enhanced minds do? Research methods on triply enhanced humans, or build AI minds operating at computer speeds. And an AI would be able to reprogram *itself,* directly, to run faster – or *smarter.* And then our crystal ball explodes, "life as we know it" is over, and everything we know goes out the window.

"Here I had tried a straightforward extrapolation of technology, and found myself precipitated over an abyss. It's a problem we face every time we consider the creation of intelligences greater than our own. When this happens, human history will have reached a kind of singularity – a place where extrapolation breaks down and new models must be applied – and the world will pass beyond our understanding."
– Vernor Vinge, *True Names and Other Dangers*, p. 47.

Way back in the dark ages of the late 1990s and the early aughts, only a handful of wild-eyed visionaries were touting this Vingean vision of the future.

My impression from interacting with Yudkowsky at that stage (mostly online, occasionally in person) was that, back then, he seemed more excited than frightened about this sort of vision. I got the feeling that, at that stage, he was hoping he could personally solve the problem of making a reliably human-friendly AGI reasonably soon, well before anyone else would make an unreliable or unfriendly AGI. When Yudkowsky led the founding of the Singularity Institute for AI in 2000, the vibe surrounding the organization seemed to be that Yudkowsky was hard at work trying to figure out the solution to the "Friendly AI problem," and might just pop up with a breakthrough any day now. Due to his background as a child prodigy, and his quick wit and obviously deep mind, he seemed to achieve an almost transhuman status in the minds of many SIAI enthusiasts of that era. I heard it said many times, in informal conversation, that Yudkowsky might be the only person alive smart enough to solve the Friendly AI problem.

As time went on, though, it was clear that Yudkowsky became more and more impressed with the difficulty of the problems involved with his quest for Friendly AI, both the issue of clearly articulating what "friendly to humans or human values" really means, and the issue of making an AGI retain its human-friendly goals and values as it becomes massively, transhumanly superintelligent via means beyond human comprehension. He also became more and more impressed with the ability of the human mind to fool itself and make basic cognitive errors, and started pondering the extent to which human rationality was really up to the task of creating AIs that would grow into reliably friendly transhuman minds.

This latter interest, in rationality, ended up occupying a lot of Yudkowsky's time during the last decade – including a fairly large percentage of his blog posts, and his popular fan-fic series Harry Potter and the Methods of Rationality (Yudkowsky 2015). It also led him to carry out a major social-engineering initiative aimed at improving the level of rationality in the Singularitarian and tech communities, which led to phenomena like "rationality boot camps" and the foundation of the Center for Applied Rationality (CFAR), a spin-off of SIAI/MIRI.

These days, Yudkowsky is still much admired within CFAR/MIRI circles, and various of his ideas have spread far and wide. There has been a large amount of interaction between SIAI/CFAR/MIRI folks and the more academic crowd at Bostrom's Future of Humanity Institute; and via this connection along with many other means, Yudkowsky's ideas have wended their way toward the mainstream. With only mild oversimplification, one might say that Bostrom's *Superintelligence* is a triumphal work synergizing Bostrom's communication and analytical ability and Yudkowsky's creative ideas. Of course, this oversimplification isn't quite fair to Bostrom, who contributed a lot of detailed ideas to his book as well. But as for the Big Ideas in *Superintelligence*, it really seems that every one was articulated at some point previously by Eliezer Yudkowsky.

Today in 2015, lurking around the edges of the MIRI/FHI world, one doesn't feel so much of a vibe that Yudkowsky is likely to singlehandedly emerge one day with a working Friendly AI, or even a detailed design for one. Rather, MIRI and FHI are seeking to grow a community of researchers who, together, will develop a research program giving concrete, rigorous form to Yudkowsky's often rather abstract and slippery ideas.

This process is already underway, in a sense – MIRI has hired a stream of excellent young mathematicians who have started proving theorems inspired by Yudkowsky's line of reasoning about Friendly AI. None of these theorems yet have any particular implications for the practical world. But as exploratory science/math, it's certainly interesting stuff.

5. Core ideas of Yudkowsky/Bostrom/MIRI/FHI–ism

As I've emphasized above, I agree with Bostrom and Yudkowsky on a number of (to me) fairly "basic" Singularitarian/transhumanist points, which nevertheless are not agreed on by the bulk of the scientific community or the tech media at the present time. In his book, though, after establishing these basic points, Bostrom goes in a number of directions that I don't wholly agree with. On essentially all of these points he concurs with Yudkowsky's previously expressed views.

The core tenets of the Yudkowsky/Bostrom/MIRI/FHI perspective, to summarize a bit crudely, go roughly like:

1. Superintelligence is very likely possible, according to known physics.[2]

2. A very likely path to superintelligence is the creation of machines that can improve themselves and increase their own intelligence, e.g. self-modifying software programs.[3]

3. In most cases, the future trajectory of a self-improving, superhuman intelligence will be very hard for human beings to predict.[4]

4. If one creates a human-level AGI with certain human-friendly goals, and allows it to self-modify freely, the odds are high that it will eventually self-modify into a condition where it no longer pursues the same goals it started out with.[5]

5. Most likely, a self-modifying superintelligence will end up pursuing goals that have little consonance with human values.

6. Quite likely this would result in the end of the human race, as a superintelligence without much consonance with human values would probably have no more reason to care about humans than we humans do to care about the ants in the dirt under a construction site where we want to erect a building.[6]

I tend to agree with points 1–4, partly with point 5, and not so much with point 6. In the overall spectrum of human attitudes, this places me very close to Bostrom and Yudkowsky – since, of course, most current humans consider point 1 highly questionable; and would consider points 2–5, if they ever thought about them, as strange and dubious science-fiction speculations. On the other hand, our point of disagreement, 6, is a fairly significant one, at least from a human perspective, since it pertains to the probability of annihilation of the human race.

One of the stronger points of Bostrom's treatment is the way he clearly distinguishes between what's best for one's personal, individual goals, and what's best for humanity as a whole.

---

[2] Bostrom does not say this in so many words, but alludes to this perspective in dozens of places. E.g. on p. 48, Bostrom notes that achieving human-level AI via whole-brain emulation shouldn't require any huge breakthroughs, reviewing details such as reconstruction of 3D neuroanatomy from electron microscope images. On p. 38, he says "human-level machine intelligence has a fairly sizeable chance of being developed by mid-century, and that it has a nontrivial chance of being developed considerably sooner or much later; that it might perhaps fairly soon thereafter result in superintelligence."

[3] See p. 179, e.g., where Bostrom says "the capacity for rapid self-improvement is just the critical property that enables a seed AI to set off an intelligence explosion."

[4] See e.g. discussion on pp. 179–183; and p. 47.

[5] Bostrom (e.g. in Chapter 7) argues that any intelligent system will tend to adopt certain intermediary goals oriented toward (crudely speaking) securing power for itself; and that superintelligences, once they have enough power, will pursue a broad variety of "final goals." On page 141, for instance, he suggests that "the first superintelligence … could well have non-anthropomorphic final goals," and the likelihood of this sort of outcome is alluded to frequently throughout the book.

[6] The "ant" metaphor is drawn from Hugo de Garis's talks on the future of AI. But this theme occurs throughout Bostrom's book, e.g. in Ch. 8 which is titled "Is the default outcome doom?"

For instance, from my personal individual perspective, it makes a big difference if a beneficent super-AGI is created in 2030 rather than 2300, because it is fairly likely I'll still be alive in 2030 to reap the benefits – whereas my persistence by 2300 is much less certain, especially if AGI doesn't emerge in the interim.

On the other hand, from the point of view of society as a whole, whether AGI comes in 2030 or 2300 may not make much difference. One could argue that a few hundred years is just a blip on the scale of human history, let alone in the history of the universe, although this isn't a fully compelling argument, given the current state of the biosphere and the overexploitation of resources due to accelerated technological change and population growth. One could also argue that humanity's near future well-being, and perhaps survival, critically depend on getting some help from very smart AGIs. AGI's existential risk may be more than balanced by the capability of AGI to help us avoid human-made disasters in the next decades or century.

Bostrom's book is concerned mainly with the humanity-wide perspective. Implicitly and sometimes explicitly, he focuses mainly on the high-level goal of maximizing the total long-term benefit of human beings (i.e. beings somehow fairly close to current humans) and maximizing the detailed control of human beings over their corner of the universe.

This goal leads naturally to Bostrom's central concern with "existential risk" – the possibility of total annihilation of the human race. If 99 per cent of the human race is killed off, then there's still a decent chance the remaining 1 per cent can revive from whatever bad situation it finds itself in, and make dramatic advances. Humanity still has potential to explore the physical universe, create AGIs and nanotech, and whatever else. But if 100 per cent of the human race is killed off, the future of humanity seems far less rosy – the only positive scenarios are those in which aliens or other unknown forces decide to resurrect the extinct human race.

With this in mind, much of Bostrom's analysis focuses on the dire existential risk that he sees AGIs as posing. He gives short shrift to existential risks posed by other technologies such as nanotech, synthetic biology and so forth, and focuses squarely on the potential risks of superintelligence. In a way this is understandable – superintelligence is a big enough topic for a book, and these other topics (e.g. especially the risk of nanotech to turn us all into "gray goo") have been discussed extensively elsewhere. On the other hand, I feel he also gives short shrift to the potential that advanced AI has to protect humanity from the risks posed by these other technologies – a potential that is directly relevant to some of his core points. I'll return to this issue below.

6. The "Scary Idea" redux

One of the most dubious of Bostrom's and Yudkowsky's ideas regards the relationship between goals and intelligences. I am referring to their emphasis on the idea referred to as

> **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. (p. 107)

This seems a strange sort of statement to preoccupy on. Among the key words here are "in principle."

In principle, according to classical thermodynamics, your head could spontaneously reassemble itself into a fully functioning robotic three-headed chicken with a passion for the writings of Chairman Mao. But these same laws of thermodynamics suggest such an event is fabulously unlikely, and we don't spend

much time worrying about it. The point is, where math and science are concerned, what is "in principle possible" is not generally of much direct interest.

According to some interpretations of the terms involved, it might be true that every level of intelligence could in principle be paired with almost any goal. But there are major gotchas. First of all, the notion of "level of intelligence" is not well-defined. The scope of possible intelligences is vast, and it's not clear that there's any single most-sensible way to rank them according to "level." Couldn't it be that some types of intelligence are particularly well suited for some sorts of goals, and other types of intelligence are particularly well suited for other sorts of goals? Couldn't this be true even for systems with a high "level" of intelligence according to some reasonable definition of "level"? It seems almost obvious that some kinds of intelligences might be REALLY BAD at figuring out how to achieve certain kinds of goals.

Also, it is not clear how this sort of general consideration relates to the class of systems realizable in our physical universe. Some intelligence/goal pairings might be physically possible but extremely unlikely to occur under reasonable assumptions, sort of like my three-headed communist chicken.

Bostrom does acknowledge some of these points, to some extent; e.g. he recognizes that a certain cognitive architecture may be required to handle complex goals, which is why his formulation of the thesis contains the hedge "more or less." But still, he considers the purported approximate truth of the orthogonality thesis more important than the potential deviations therefrom that he admits may occur.

The related question I care more about, though, is: In practice, which goals are *likely* to be allied with which kinds and levels of intelligence, in reality? What goals will very, very smart minds, existing in the actual universe rather than the domains of abstract mathematics and philosophy, be most likely to aim for?

About this more practical question, Bostrom's main suggestion is (to paraphrase rather loosely) that superintelligences will likely be concerned with self-preservation and with propagating themselves throughout the universe. He does acknowledge that, in some cases, superintelligences might end up favoring their own destruction. But he argues they will generally focus on the integrity of their goal content and on the acquisition of relevant resources – which in practice will usually amount to the same thing as preserving and propagating themselves and their influence. In this regard, Bostrom's arguments are essentially the same ones given a few years earlier in Steve Omohundro's classic paper "Basic AI Drives" (Omohundro 2008), which generated a great deal of excitement in the Singularitarian community. But, much as I admire and respect the minds of Steve Omohundro and Nick Bostrom, I detect more than a small hint of anthropomorphism and biomorphism in this line of thinking.

Bostrom, in his usual style, presents these ideas with a great number of hedges, such that one can't really argue with the details of what he says:

> **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 a broad spectrum of situated intelligent agents. (p. 109)

OK … well … that's hard to argue with, because it's not clear what a "wide range" and "broad spectrum" actually mean. 10 per cent? 90 per cent? 1 per cent? .01 per cent? The problem is that neither Bostrom nor anyone actually knows enough to put a number on this, even a rough one. Actually, what percentage of physically-reasonably-likely intelligences will care about these human-like "instrumental values" –

which, while phrased by Bostrom in relatively abstract terms, appear to me very similar to classic human values like survival, self-protection, power-seeking and so forth? We don't know. We can say it will be a "broad spectrum," but that doesn't tell us much of anything. This just tells us that there may be some physically-feasible intelligent agents who aren't that much like humans but still share these general evolution-oriented values. But there might be a lot MORE, or a lot more likely, non-human physically-feasible intelligent agents who have other values entirely.

This is typical of Bostrom's rhetorical style in the book: He is very careful not to make statements that could be definitively proved wrong. But on the other hand, his carefully couched and hedged statements remain highly *suggestive*, so that it seems very clear what he actually thinks, behind all the painstaking academic formulations.

To me, the "Instrumental Convergence Thesis" – in spite of the fancy name – feels a bit like a dog projecting that the main goals of humans and all future super-dog intelligences will be sex, meat, and running around in the fields How can we really presume to know what will guide the activity of superhuman superminds? We have evolved as a population of individual minds in individual bodies, struggling at times for resources – but as Bostrom himself points out, future superintelligences are fairly likely to fuse in some way into some sort of fairly-unified, distributed mind-network. What kind of motivations will such a mind-network possess? Perhaps motivations significantly beyond maintenance of egoic boundaries and accumulation of resources? Abundance, of the sort that advanced nanotech and femtotech and so forth will enable, will shift the motivational systems of all intelligences capable of using the tech, in ways that are hard to currently foresee.

Omohundro, in his "Basic AI Drives" paper, argues (among other things) that systems with the drive for self-preservation will be more likely to survive – thus he argues that self-preservation is an instrumental value that almost any superintelligence will derive as a subgoal of its final goals. At first this seems obvious, then after more reflection it starts to seem very dubious. It may well be that "self-preservation" is an anthropomorphic or biomorphic idea, and very advanced AGI systems might go far beyond such notions.

In a scenario of competition among powerful AGIs, victory might well go to AGI systems that were free to create new, more powerful AGI systems very DIFFERENT from themselves, without worrying about whether these new systems had a continuous identity with their predecessors. If system A can create any new system A1 that it dreams of, but system B can only create a new system B1 that has continuity of identity with B – then which one has more liberty in designing a new system able to win a competition between AGI systems? Non-self-attached systems would most likely do better at spawning new systems with the capability of winning.

Similarly, the desire to maximize a fixed utility function – and maintain this same utility function throughout successive self-modifications – would likely make an AGI system less powerful than potential competitors. So, one point is that, generally speaking, in a struggle for power among powerful, superintelligent AGI systems, one might expect victory to go to systems that didn't care about continuity of identity or utility function preservation, but were simply concerned with creating the most powerful successor systems. Bostrom acknowledges this sort of possibility, in a vague way and almost in passing.

But he doesn't go very far down the road of exploring what sort of universe might eventuate if thought-patterns not tightly connected to identities or goals in human-like senses became the dominant factor. Is it indeed likely that such a universe would be hostile to traditional human beings who wanted to retain the traditional human form? If so, why? Here, as in most of Bostrom's book, the distinction between "danger to traditional humans is possible" and "doom for traditional humans is the default scenario" is not carefully drawn.

Omohundro pays a lot of attention to the goals he thinks an AGI would need to have in order to survive in a competitive environment where AGIs are vying for resources. I think his analysis is interesting but too simplistic, because it assumes that the AGIs involved are vying for resources for themSELVES, so that they can pursue maximization of their utility functions – when in fact, in a post-Singularity context, self-focused, explicitly-utility-function-maximizing AGI systems may be among the less powerful and intelligent ones out there (if they exist at all).

But since I don't think a simple competitive scenario between powerful AGIs is terribly likely, I am not impressed by the argument for humanly "obvious" values like self-preservation and goal-preservation to be "instrumental" and universal across AGI systems. And I think Bostrom's more abstract version of the argument is basically just as problematic as Omohundro's more direct version.

A more important, related point is: What properties a "broad spectrum" of possible intelligences may have may not even be a relevant issue to think about, apart from its pure intellectual interest. The actual AGI minds we will see in our future are not going to be randomly selected from the space of all physically-possible intelligent systems – they are going to be conditioned by their interactions with the physical universe and the human world. They are going to grow up from a specific starting-point – the AGI systems we create and utilize and teach and interact with – and then they are going to be shaped, as they self-modify, by their interactions with our region of the physical cosmos. Their experiences will likely lead them in directions very different from anything we currently expect – but in any case, this sort of process is nothing like choosing a random mind from the space of all possible minds. So the fact that Bostrom can conceive of a superintelligence being paired with any arbitrary goal he comes up with (the orthogonality thesis) doesn't really matter much. And the fact that Bostrom can conceive of a "broad spectrum" of AIs that converge to certain instrumental goals (the instrumental convergence thesis) doesn't matter much either – if the AI minds actually created by humans live in a different broad spectrum within the even broader spectrum of possible minds.

In the portions of his book dealing with these issues, Bostrom is basically presenting his own analytic-philosophy-esque version of what I called, in a 2010 essay, "The Singularity Institute's Scary Idea" (Goertzel 2010). Bostrom presents the "Scary Idea" more rigorously and soberly than anyone involved directly with SIAI/MIRI has so far, but ultimately his arguments for it aren't any stronger than theirs. The basic gist remains the same:

1. There is a wide variety of possible minds, many of which would be much more powerful than humans and also indifferent to humans.

2. It is [possible/highly probable] that the highly powerful AGIs we humans create, and their descendants, will be indifferent to humans.

3. Advanced AGIs that are indifferent to humans and human values will likely do bad things to humans and do bad things from a human-values perspective.

Point 1 seems hard to argue with. Yes, human-like or human-friendly minds would seem to constitute a small fraction of the kinds of minds that are physically possible.

Point 3 is hard to know about. Maybe AGIs that are sufficiently more advanced than humans will find some alternative playground that we humans can't detect, and go there and leave us alone. We just can't know, any more than ants can predict the odds that a human civilization, when moving onto a new continent, will destroy the ant colonies present there.

Point 2 is the core. What can be rationally argued is that the outcome described is *possible*. It's hard to rule out, since we're talking about radically new technologies and unprecedented situations. However, what prompted my "Scary Idea" essay was the frequency and vehemence with which various SIAI staff and associated people argued that this outcome is not only possible but *highly probable*, or even *almost certain.* These SIAI folks liked to talk about how a "random mind," plucked from mind-space, almost surely wouldn't care about humans. Quite possibly not. But an artificial mind engineered and taught by humans is not random, and its descendants are not random either.

There are certainly possibilities here that are scary, from a human-values perspective. The uncertainty about how likely these possibilities are, can certainly feel disturbing – it even does to me, at times, and I'm a tremendous AGI optimist. But this uncertainty doesn't merit assignment of a high probability to the outcome.

In Chapter 8 of his book, Bostrom considers various particular aspects of the potential danger from superintelligence. He discusses the "treacherous turn" problem, i.e. the possibility that a young AGI that seems benevolent might then turn destructive in ways its early human observers did not predict:

> **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. (p. 143)

He considers various subtle ways in which this might occur. For instance, perhaps when the AI gets smart enough, it is able to figure out whole new ways of achieving its goals – which are nastier, from a human view, than the methodologies it found when it was less intelligent.

Of course, in principle, there is no way to rule out this kind of treacherous-turn outcome. But there is also no reason to expect it to be especially likely. This would also seem to be the sort of possibility that could be explored via simulations and experiments with AIs at infrahuman levels of general intelligence. Under what conditions do intelligent agents tend to display treacherous turns? What does the degree of coupling, and the quality of the relationship, between an intelligence and its creators have to do with the probability of a treacherous turn occurring?

The "perverse instantiation" problem also gets attention, e.g. (pp. 145–147):

> **Final goal**: "Make us smile"
> **Perverse instantiation**: Paralyze human facial musculatures into constant beaming smiles.
>
> **Final goal**: "Make us smile without directly interfering with our facial muscles"
> **Perverse instantiation**: Stimulate the part of the motor cortex that controls our facial musculature in such a way as to produce constant beaming smiles.
>
> **Final goal**: "Make us happy"
> **Perverse instantiation**: Implant electrodes into the pleasure centers of our brains.
>
> **Final goal**: "Act so as to avoid the pangs of bad conscience"
> **Perverse instantiation**: Extirpate the cognitive module that produces guilt feelings.
>
> **Final goal**: "Maximize the time-discounted integral of your future reward signal"
> **Perverse instantiation**: Short-circuit the reward pathway and clamp the reward signal to its maximal strength.

The key point here is that, as Yudkowsky argues elegantly in Chapter 279 of *Rationality: AI to Zombies*, "value is fragile." Pithy summaries of complex human values evoke their commonly accepted meanings only within the human cultural context.

For AI researchers not enamored of the reinforcement learning approach, though, this may seem a problem particular to AI methods centered on providing AIs with explicit utility functions. What if an AI isn't conceived as a utility function maximizer, but instead as a different sort of complex system, which engages with humans and implicitly absorbs complex human values from its interactions? Of course, the AI could still learn the wrong lessons; or it could take a treacherous turn. But by and large, the "perverse instantiation" problem as formulated seems a refutation of certain simplistic approaches to AGI architecture and instruction, rather than a general point about machine intelligence or superintelligence.

In accordance with Bostrom's carefully guarded style, there is no sentence in his book in which he comes right out and says it's probable that powerful AGIs created via an "ordinary" process of human science and engineering would kill all humans and act radically against human values. But to me, at least, this undertone is clearly there throughout wide swaths of the book. It's there in his choice of which topics to emphasize and which ones to downplay, and it's in his offhanded phraseology here and there.

Many of Bostrom's hints are not especially subtle; e.g. the title of Chapter 8 is "Is the default outcome doom?" The answer given in the chapter is basically "maybe – we can't rule it out; and here are some various ways doom might happen." But the chapter isn't titled "Is doom a plausible outcome?", even though this is basically what the chapter argues. Also, there is no corresponding chapter "Is the default outcome utopia?", although certainly arguments could be marshalled in favor of this alternative hypothesis, with as much rationality and color as Bostrom's intimations of doom. There's nothing wrong with writing a polemic arguing one side of an argument; but what we have here is a special kind of polemic – a polemic that eschews extreme language in favor of a verbal tone and impression of great care and balance, but presents a very one-sided polemical view nonetheless.

For another (key) example, consider a passage like this, exploring the possibility of a "fast takeoff" in which superintelligence emerges rapidly from early-stage AGI:

> A fast takeoff occurs over some short temporal interval, such as minutes, hours or days. Fast takeoff scenarios offer scant opportunities for humans to deliberate. Nobody even notices anything unusual before the game is already lost. In a fast takeoff scenario, humanity's fate essentially depends on preparations previously put into place. At the slowest end of the fast takeoff scenario range, some simple human actions might be possible, analogous to flicking open the "nuclear suitcase," but any such action would either be elementary or would have been planned and preprogrammed in advance. (p. 252)

This is an extremely slanted way of writing about "fast takeoff" scenarios, is it not? The assumption that the AGI resulting from the fast takeoff will be bad for humans is baked into the prose. The possibilities that the resultant AGI might lead the way toward an amazingly positive future for humans, or might just disappear into space or some other dimension and ignore humanity, are not even mentioned and thus utterly downplayed. What we find in *Superintelligence* are careful philosophical formulations arguing why terrible outcomes are *possible*, and then more practical discussions predicated on a "plan for the worst" sort of attitude, and sweeping aside positive possibilities.

When I read the above passage, I thought "What does he mean, *the game is lost*?…" What is the game? Why do we want to view the process of creating the next generation of minds as some sort of game of Us versus Them? Why is this way of looking at the future being implicitly presupposed? A similar

presupposition is made throughout the book, between the lines of the careful analysis, and above all in the choice of which topics to focus on.

The potential of AGI to help humanity think better about the complex problems facing us – including problems related to the future of technology and society as a whole – gets short shrift in Bostrom's book. So does the potential of AGI to alleviate suffering, and the potential of AGI to help decrease the risks (existential and otherwise) of other advanced technologies. These points are not denied nor totally ignored, but they are quickly brushed past, whereas fairly wild speculations that provide more emotional support to the Bostrom/Yudkowsky perspective get a lot more space. This strategy works, in part, because it plays into aspects of human psychology that give more focus to negative scenarios than positive ones, in certain contexts (which is obviously related to the prevalence of scare scenarios in SF movies about AI and robots, as well). Nothing riles people up and gets them excited and feeling all communal like a common enemy – even if the common enemy is a purely hypothetical superintelligence!

Bostrom's excellent book should be understood as a master debater presenting his best arguments for his own perspective, rather than as a genuine attempt at a balanced treatment of the subject. For what it is, *Superintelligence* is an excellent book. I would never recommend it as the ONLY book for a newbie to the area to read about the future of AGI – it's too narrow in its outlook, too thoroughgoing in its precautionary focus. But I would definitely include it on a "Future of AGI reading list" of five to ten books – it presents its perspective very, very well.

7. AGI in the mix

Bostrom's hasty, somewhat cavalier treatment of the likely interplay of different advanced technologies during the next decades and centuries merits further emphasis. To me this is a very real issue. As technology advances, fewer and fewer people with less and less specialized knowledge are able to cause more and more damage to more and more people.

In spite of the much-publicized increase in surveillance via government agencies, the capability of governments – or anyone else – to deal with the threats posed by terrorist groups or crazed rogues armed with advanced technologies is questionable and certainly far from clear. The video and audio data gathered by intelligence agencies from various sources is far more than they can scan through usefully given current technologies, and it's not clear how far narrow-AI is going to take us in this regard.

So far, the power of biological weapons has been limited due to difficulties in distributions. Biotoxins tend to dissipate in the wind, rain, and sun. Will these problems be overcome as genomics, proteomics, and metabolomics are solved, and synthetic biology advances? As biotech and nanotech advance together, what will be the relative rate of advance of defensive versus offensive technology?

There are massive unknowns here, but it doesn't seem sensible to simply assume that, for all these non-superintelligence threats, defenses will outpace offenses. It feels to me like Bostrom – in his choices of what to pay attention to, and his various phrasings throughout the book – downplays the risks of other advanced technologies and over-emphasizes the potential risks of AGI. Actually there are massive unknowns all around, and the hypothesis that advanced AGI may save humanity from risks posed by bad people making dangerous use of other technologies is much more plausible than Bostrom makes it seem.

Bostrom does make a good point that many other technologies require a medium-term build-up of hardware infrastructure, whereas AI has a greater likelihood of being created in stealth (e.g. via some genius hackers writing code in their living rooms, etc.). It would be much more shocking if someone produced a molecular assembler than an AGI next year, simply because one would expect a molecular assembler to require some hardware advances that haven't yet been seen, and big hardware advances tend

to be more noticeable. On the other hand, the gradual build-up of infrastructure for synthetic biology is already happening, right in the open – so I wouldn't be tremendously shocked if some obscure lab in Eastern Europe or Asia came out next year with a radical advance in the ability to engineer new organisms, perhaps including unprecedentedly robust pathogens.

Giving a high subjective estimate of the odds that AGIs will kill everyone, and a low subjective estimate of the odds that other technologies will pose dire dangers that AGI might be able to squash, leads to Bostrom's conclusion that we'd do best to slow down AGI development. But, while Bostrom is certainly welcome to his own intuitive subjective probability estimates, he doesn't give particularly strong arguments for them. Time and time again, his prose wends from "possible" and its analogues to "quite likely" and its analogues without much justification. My own subjective estimates happen to be otherwise – I think the odds of advanced AGI killing everyone are not that high, and the odds of AGI saving us from dangers posed by terrorists wielding other advanced tech are pretty high. But debating subjective probability estimates regarding radically new, substantially unpredictable future scenarios is a tricky matter.

One might argue that, if it's all so uncertain, maybe we should just stop developing ALL these technologies till we know what's going on. That was basically Bill Joy's argument in his classic turn-of-the-millennium essay "Why the Future Doesn't Need Us" (Joy 2000). There's a certain appealing coherence to this view, but the counterarguments are pretty clear. First, the human race has NEVER known what was going on. Ever since we emerged from our hunter-gatherer existence to form "civilized" society, we've been plowing on into unknown territory, taking the risks in stride and embracing radical novelty. Second, there seems no practical way for any of the readers of *Superintelligence* or Bill Joy's essay to stop the development of all advanced technologies, even if (like, say, the Unabomber did) they wanted to. These technologies are being developed around the world by a wide variety of actors for their own reasons and due to their own interests.

Stopping everything is not feasible short of imposing a global police state of some sort, which would be very difficult, highly ethically questionable, and probably wouldn't work anyway. Stopping just AGI or (say) just nanotech and continuing everything else would have quite dubious value, as in the overall context of advanced tech development, each individual technology has both dangers and defensive potential.

In particular, I have argued that AGI has a strong potential to play a temporary role as a "nanny" technology (Goertzel 2012), protecting humanity from its own tendencies toward technological self-destruction, while more and more advanced AGI emerges in a relatively measured way. Arguably, such an "AGI Nanny" could manage the transition to and through a Singularity much better than human beings or purely human institutions. An AGI Nanny would in a sense be a kind of "police state" – but if policing of dangerous technologies was combined with compassionate ethics, abolition of human-scale material scarcity, and coherent rational pursuit of a positive Singularity, one would have something very different from the "police states" that have existed in human history. This "AGI Nanny" argument is certainly not bulletproof and conclusive, but I personally feel it has significantly fewer and smaller gaping holes than Bostrom's key arguments in *Superintelligence.* Indeed some sort of AGI Nanny type technology may emerge semi-spontaneously via the emergence of AI-powered sousveillance technologies across the Internet, motivated by a constellation of other goals on the part of various groups.

8. Friendly AI vanguardism

Another current of thinking that has often been quite explicit in SIAI/MIRI writing, and is present in Bostrom's book, though in a subtler form, is what I think of as "Friendly AI vanguardism." Put roughly, what I mean here is the idea that the problem of building a Friendly AI – a superintelligence that won't

kill everybody – is so hard that only a small, elect crew of extraordinarily rational and intelligent people with the right goals and mindset can possibly be trusted to work on it.

This attitude has been highly evident in the SIAI/MIRI community since its beginning. The informal plan of SIAI in its early years, from what I could tell, appeared to be to gather a group of brilliant technologists and scientists to work together in secret on creating a Friendly AI which would then save the world. SIAI/MIRI has matured a great deal since then, yet much of this vibe still remains.

And this vibe resonates closely with the explicitly elitist attitude promoted by Peter Thiel, who was a major SIAI donor for a number of years (Reinhart 2011). Thiel has, for instance, spoken very harshly about the modern education system, and he funded a handful of bright kids less than 20 years old to drop out of college (the "Thiel Fellowship"). But he doesn't seem to have much understanding of the transformative role that college can play for people from an underprivileged background, e.g. people in the developing world; and even among the privileged youth who are its target, the Fellowship has had dubious effectiveness (Lawrence 2013). Along the same conceptual lines, both Thiel and SIAI/MIRI have been very interested in Math Olympiad winners, chess champions, and so forth (personal communication). There is an attitude of focusing attention on maximally leveraging the ability of the people society has identified as the "best of the best."

Bostrom does not come right out and say that he thinks the best path forward is for some small vanguard of elite super-programmers and uber-scientists to create Friendly AGI, but he almost does. In a passage that comes off a bit scary-sounding to me, he notes the conditions under which he thinks a government/corporate funded AGI effort could perhaps rationally be allowed to proceed:

> [A]n international project to develop safe superintelligence would … have to be constituted not as an open academic collaboration but as an extremely tightly controlled joint enterprise. Perhaps the scientists involved would have to be physically isolated and prevented from communicating with the rest of the world for the duration of the project, except through a single carefully vetted communication channel. The required level of security might be nearly unattainable at present, but advances in lie detection and surveillance technology could make it feasible later this century. (p. 253)

And then he notes the possibility of the UN allowing a small group (maybe just one person) to proceed with AGI R&D on approved lines, under their wing:

> [B]road collaboration does not necessarily mean that large numbers of researchers would be involved in the project; it simply means that many people would have a say in the project's aims. In principle, a project could involve a maximally broad collaboration comprising all of humanity as sponsors (represented, say, by the General Assembly of the United Nations), yet employ only a single scientist to carry out the work. (p. 253)

What he is advocating here, in his dry professorial style, is actually something quite dramatic: For the UN and all the governments of the world to come together to control AGI research and development, protecting and fostering an elite AGI R&D effort carried out under their auspices by a small group, potentially even just by one person. (And who is that one person going to be? Eliezer Yudkowsky perhaps?)

Underlying these somewhat extreme proposals is the vibe that the "control problem" or "Friendly AI problem" is something that can more likely be solved via sustained hard thinking by a small, select team of super-smart humans, than via ordinary human processes of science or engineering. As Bostrom says, in the context of considering the impact of brain enhancement on superintelligence,

> One reason why cognitive enhancement might cause more progress to have been made on the control problem by the time the intelligence explosion occurs is that progress on the control problem may be especially contingent on extreme levels of intellectual performance – even more so than the kind of work necessary to create machine intelligence. The role for trial and error and accumulation of experimental results seems quite limited in relation to the control problem, whereas experiential learning will probably play a large role in the development of artificial intelligence or whole brain emulation. (p. 236)

I read these passages and thought – Whoa!! What a huge win for the Yudkowsky/SIAI folks, to have their ideas written up and marketed in a way that managed to appeal to Bill Gates, Elon Musk, Stephen Hawking and so forth! And indeed the win was real financially as well as media-wise: Bostrom's book helped entice Elon Musk to donate $10M to the Future of Life Institute for research into how to make AI beneficial rather than destructive, and the two biggest chunks of the FLI funding went to Bostrom's Future of Humanity Institute (FHI) and Yudkowsky's MIRI (in that order).

One core idea here seems to be that a few brilliant, right-thinking mathematicians and philosophers locked in a basement are most probably our best hope to save humanity from the unwitting creation of Unfriendly AI by teams of ambitious but not-quite-smart-enough AI developers. This idea was batted around frequently on the SL4 futurist email list (operated by Eliezer Yudkowsky) in the late 90s and early aughts, but that was an in-crowd of Singularitarian geeks. Now the meme has hit the big time.

Bostrom and Yudkowsky, from what I have seen, are ethical and peaceable individuals who genuinely want the best for all mankind, as well as wanting to see amazing transhuman possibilities come about in a safe and well-managed way. However, their tendency toward elitism does not strike everyone as benign in its potential consequences. Hruy Tsegaye, an Ethiopian writer and educational technology developer who has written about the implications of advanced technology for Africa (Tsegaye 2015), had this to say about Bostrom's scenario of governmentally-enforced restriction of AGI development to a chosen few:

> This is the highway to tyranny. The current world is stained with odious inequality because of such attitudes and systems. The "few" will be in control of this game-changing hi-tech and then who will control these few? Obviously those who have guns and money will control these few. And then what? Oh and then instead of protecting mankind from AGIs we have AGIs destroying the majority of mankind on behalf of the "few." (personal communication)

Bostrom appears to place slightly more faith than Yudkowsky – and vastly more than Tsegaye – in the potential of the US or other governmental bodies to regulate AGI development. On this my intuition also veers to the skeptical side. Regulating nuclear weapons development seems to strain the global political system, and this seems a much easier task than regulating AGI development: it requires specialized materials and expensive machinery, and it doesn't have immediate and obvious practical and commercial benefits besides warfare. But the commonality of a vanguardist perspective, which these two thinkers share, is what I want to point out here.

This particular aspect of Bostrom and Yudkowsky's thinking is highly relevant to my own practical work, as my own AGI efforts are centered on an open-source AGI project, OpenCog. Developing AGI in the open source domain is precisely the opposite of the vanguarding perspective that Bostrom and Yudkowsky advocate. The potential risks of the open-source perspective are evident: people with bad aims could take your open-source AGI code, privately fork it, and use it to develop killer robots or evil superintelligence. On the other hand, the potential benefits are also evident: One gets the creative, technical, moral, and social insight of the whole world brought to bear on one's AGI project, not just the thinking and intuition of a small, self-selected elite group. My OpenCog collaborator, Joel Pitt, and I

explored these themes in a preliminary way, in a 2012 essay titled "Nine Ways to Bias Open-Source AGI Toward Friendliness" (Goertzel and Pitt 2012).

At the present time, it seems very unlikely that Bostrom's vision of a tiny, UN-sanctioned elite Friendly AI group, sheltered from competition by government regulation, is actually going to come to pass. Rather, what seems to be happening in the real world is that large companies are jumping into the AGI game. AI progress is not being driven by MIRI protected by the UN, but rather by Google, Baidu, Facebook and IBM. Of course, the work these companies are doing is not yet producing human-level thinking machines, but arguably the application-focused "narrow AI" work they are doing is contributing more to the path toward AGI than FHI's philosophizing and MIRI's theorem-proving and online-community-building. In the current scene, OpenCog is something of a maverick attempt to guide the center of AGI development away from the mega-corporations and toward the more diffuse network of open-source volunteers, university researchers and students, and startups leveraging and contributing to open source projects as part of their own development.

The open source approach also would seem to increase the probability that multiple AGIs will emerge in different places for different purposes, creating a balanced field of multiple AGIs that will have to coordinate with each other, perhaps fusing into some sort of mindplex. Whether this wards off disasters better than having a single AGI project is open to debate, though my guess is that it does.

Open-source AGI is a large topic on its own and beyond the scope of this review article; I bring it up here mainly to point out that the FHI/MIRI vanguardist perspective is far from the only viable view on how the future of AGI development should proceed. There are a lot of possibilities out there, and the ones that Bostrom focuses on are not particularly among the more realistic, nor necessarily the most desirable.

9. Open-ended intelligence

When one hears smart, hard-driving, committed people talk about a certain set of issues in a certain way, it's easy to get caught up in their way of framing the issues. If one doesn't agree with their perspective, one can get sucked into debating the issues on their terms. This can then direct attention away from other, fundamentally different perspectives on the same issues, which may have their own considerable value.

I see some risk of this happening now with the topic of superintelligence – which is one reason I've taken the time to write this essay. Bostrom, Yudkowsky, Musk, Hawking and their ilk definitely deserve to have their opinions heard. But this doesn't mean we need to accept their framing of the issues surrounding superintelligence.

An utterly different framing, in my view more deeply grounded in profound thinking about humanity, intelligence and the universe, is provided by David Weinbaum (Weaver)'s notion of "open-ended intelligence" (Weinbaum and Veitas 2015). As Weaver and his colleague Viktoras Veitas put it,

> **Open-ended Intelligence** is a process where a distributed population of interacting heterogeneous agents achieves progressively higher levels of coordination. In coordination here we mean the local resolution of disparities by means of reciprocal determination that brings forth new individuals in the form of integrated groups of agents (assemblages) that exchange meaningful information and spontaneously differentiate (dynamically and structurally) from their surrounding milieu. This kind of intelligence is truly general in the sense that it is not directed or limited by an a priori given goal or challenge. Moreover, it is intrinsically and indefinitely scalable, at least from a theoretical point of view. We see open-ended intelligence manifesting all around us and at many scales; primarily in the evolution of life, in the phylogenetic and ontogenetic organization of brains, in life-long cognitive development and sense-making and in

the self-organization of complex systems from slime molds, fungi, and beehives to human sociotechnological entities.

The theory of open-ended intelligence rejects the idea that real-world intelligent systems are fundamentally based on goals, rewards, or utility functions. It perceives these as sometimes-useful, but limited and ultimately somewhat sterile, descriptors of some aspects of what some intelligent systems do in some circumstances.

What, after all, are the goals of any real-world human being? They are ideas that the human holds in their mind, or that others pose for that human. They shift constantly over time, and have only limited influence over that human's actual behavior. Similarly – but even more so – for any real-world human being's "utility function." Economists have struggled for a long time to model actual human behavior in terms of the math of utility functions – and that's just in the domain of economics, where one would think it would be most applicable. Forget about modeling human behavior in art, science, or romance in terms of utility functions. It just isn't a natural model of how we work. There is quite a substantial body of evidence that humans do not operate even closely to the rational agent model. It seems intuitively clear that rationality, while important, does not cover everything intelligence is about.

Sure, mathematically, it might be possible to model what ANY finite system does in terms of some formal utility function. But such a pursuit quickly becomes reminiscent of Ptolemaic epicycles. It's not a natural way of modeling humans, and there's no evidence currently that it's a useful way.

But even if goals and utility functions aren't a good way of modeling people, couldn't they still be a good way of modeling future AGI systems – such as superintelligences? There is a certain point here. OpenCog, for example, is one among many AI systems explicitly architected to be driven by specific, formally-articulated goals. OpenCog is far more goal-driven than any human being, far more goal-oriented in its architecture and dynamics.

However, the dynamics of a real OpenCog system embedded in a complex real-world environment are still only *guided by* the system's formal goals – not rigorously driven by them. There are also other, ambient, non-goal-driven dynamics in the system – and these can play a role in the system's ongoing redefinition of its own goals. Ultimately, the initial goals assigned to an OpenCog system by its programmers may serve as nothing more (or less) than a way of helping the system to get its own, open-ended, self-organizing intelligence off the ground.

Once an intelligent system becomes vastly smarter than humans, the odds seem to me very low that its intelligence will be well-characterized by human concepts like goals and utility functions. The paradigm of "intelligence as optimization" that seems so natural to us now may seem absurdly, quaintly limiting to our future AGI descendants, much as we would now view a dog-level model of intelligence as "getting lots of meat, having sex and running around a lot." Indeed, this dog-level view still has some relevance to human life and human society, but in important ways we've also gone beyond it – and dogs and humans are ultimately pretty similar. Bostrom, Yudkowsky, and I agree that AGIs will likely end up going far further beyond humans than we have gone beyond our vaguely dog-like mammal ancestors. But I differ from them in suspecting that these advances will also bring us beyond the whole paradigm of optimization.

Back in my first book *The Structure of Intelligence* (Goertzel 1993) I described intelligence as "The ability to achieve complex goals in complex environments." I even posited some (rather broad, with lots of free parameters and no useful calculations) specific mathematics for assessing a system's degree of intelligence according to this measure. So I can very well understand the general line of thinking Bostrom and Yudkowsky are pursuing, in modeling intelligence in terms of goal-seeking. However, my thinking

on the topic has evolved since then, and I think Bostrom's and Yudkowsky's thinking needs to evolve too.

I still consider the conception of intelligence in terms of the ability to achieve complex goals in complex environments as a useful working definition for driving progress forward in building real-world AGI systems. When I have my practical AGI system developer hat on, I'm thinking a lot about achieving complex goals in complex environments – and about exactly what these goals and environments should be.

However, I am currently a lot less convinced that this sort of perspective on intelligence is going to seem relevant to my AGI descendants 1000 years from now. If I had to pick a human concept likely to seem more relevant then, it would be what I sometimes refer to as SCADS, or Self-Organizing Complex Adaptive Systems.

Given this broader perspective, one possible rhetorical move would be to posit that "intelligence" itself is going to seem a less interesting concept to future superintelligent AGI systems. Another possible rhetorical move would be to redefine intelligence as SCADS, leaving aside "complex goals in complex environments" as a limited subclass of true SCADS/intelligence. It seems to me Weaver's strategy is essentially the second one: his notion of "open-ended intelligence" feels to me like a way of fleshing out and semi-formalizing SCADS.

Interestingly, while Bostrom is an analytic philosopher through and through, Weaver draws more of his inspiration from Continental philosophy (as well as from mathematics and science). Deleuze comes up often in his thinking.

Continental philosophy has been mocked by many in the scientific community. But I have found that, once one surmounts the initial difficulties posed by the opaque writing style customary in the field, one finds a variety of interesting insights. One must, however, let go of the notion that one is supposed to be finding definite conclusions, and knock-down arguments in favor of these conclusions, in the style that one finds in analytic philosophy. Rather, in Continental philosophy one is exploring issues, without ever fully exhausting or understanding them; one is following questions where they lead, which is usually to other questions, and then exploring the pattern of (new and old) questions that each question has led to. Analytic philosophy aims to replace imprecise intuitive concepts and arguments with exact formal analogues; Continental philosophy is more concerned with unraveling the patterns via which intuitive concepts and arguments relate to each other. Formalizations are not viewed as a preferred domain of discourse, but as one collection of relationships among many.

In this spirit, Weaver's vision of intelligence is fundamentally non-reductive. He does not aim to reduce intelligence to "utility functions plus optimization" or any other set of underlying core concepts. Rather, he views intelligence as a process of ongoing growth that necessarily goes beyond any reductive understanding one might try to impose – and also goes beyond the specific formulation in his paper, which he views as an approximation of a greater evolving understanding.

Now, reduction is not always a bad process. "Reducing" cells to molecules is helping biology a lot, as the reduction of atoms to particles has helped physics. However, reduction is not the only powerful heuristic for understanding. Evolution is best understood, it seems, via a combination of reduction-oriented genetics, and systems-oriented ecological thinking. Intelligence, at the human scale, is partially well understood in terms of goals and optimization; but the understanding becomes richer if one also takes a systems view and looks at intelligent systems as complex self-organizing processes coupled with other complex self-organizing processes, in various emergent dynamics.

An open-ended intelligence strives to achieve various goals which it considers important in various contexts, but it also re-factors its own goals, and creates new goals on the fly. It may give up a long-cherished goal seemingly all of a sudden. It may take part in newly formed emergent patterns between itself and its environment (at times, even redefining the very boundaries between the system and its environment), which give rise to whole new systems extending beyond (or including only part of) the previously existing "intelligent system."

This sort of open-ended intelligence and growth is how human intelligence evolved. I would rather not see the future development of human intelligence emulate the process of evolution by natural selection – by my human ethical standards, evolution is extremely cruel and wasteful. But I would like to see the future development of intelligence emulate the openness of the process of evolution by natural selection – and indeed I don't really think there's much of a choice. However much we try to constrain intelligence, to close off its routes of developmental possibility, it is going to open them up anyway.

What's the practical upshot of all this open-ended philosophy? We shouldn't be thinking in terms of "How can we craft an AGI system that will be provably guaranteed to maintain its initial goal system for all eternity?" Goal systems are just a heuristic guide to the behavior of any real-world intelligence anyway, and future AGIs will likely be even less adequately concisely describable in terms of goals and optimization.

Rather, we should be thinking in terms of "How can we effectively participate in the open-ended growth and development process of the next phases of intelligence, so that as these phases unfold, we can effectively manifest what we are, and genuinely feel we are part of what is to come?"

This is a different framing of the issue than the one Bostrom, Yudkowsky and kin present. However, it's actually fairly close to the framing that Ray Kurzweil provides (Kurzweil 2005). Kurzweil generally prefers to focus on the potential of the Singularity to bring improvement to everyday human life – by ending aging, by creating material wealth via nanotechnology, and so forth. But when pressed about the fundamental nature of the changes to come, Kurzweil always notes that it's not going to be "us versus them" – rather, advanced technology will become part of us, so that it's us who are evolving into massively superhuman superintelligences.

Evolution by natural selection, as an example of open-ended evolution, provides examples of fairly sudden catastrophic change (e.g. replacement of dinosaurs by mammals) and also of progressive gradual change (early primates into ape-like creatures into humans, etc.). We may well have an aesthetic and moral taste for the transition from human to posthuman intelligence to be more of a gradual transition (i.e. "gradual" morphologically even if fairly rapid temporally on the human scale). We may even be able to guide the coming transition to increase the odds this is the case. But if we do so, it won't be by creating superintelligent AGIs that rigorously obey their human-coded goals even as they interact with humanly inconceivable aspects of the environment, and modify and improve themselves in humanly inconceivable ways. The open-ended intelligence of the universe really doesn't seem to work that way.

What does Weaver himself say about Bostrom's perspective on the future of humanity? Not surprisingly, his views agree more closely than mine, though he frames them perhaps even more strongly. When I asked him about the topic, after his presentation on open-ended intelligence at the AGI-15 conference in Berlin, he noted that

> A sane path towards the development of human level and beyond AGIs, was a significant aspect of my motivation in developing open-ended intelligence as an alternative to the current thinking. I meant to say a few words on that at the conference but time was too short. In the longer version of my presentation I planned to address Bostrom's argument of orthogonality,

> that intelligence and value/purpose can develop independently. I think this is a deeply flawed understanding of the nature of intelligence. Indeed we can "rape" future intelligent agents to follow certain rigidly defined goals. Reducing intelligent agents to goal-oriented machines made to be used only as tools and motivated by externally imposed value systems is a distorted projection of human nature. Humans already did that much in slave societies. There is no happy end to such stories. I do agree that this is a very dangerous path but the danger is not in the technology as such but rather in a narrow minded and distorted approach towards the science of intelligence... The major point here is that competence and goal co-evolve and co-determine each other and cannot seriously be decoupled. (personal communication 2015)

10. Yudkowsky's resistance to the organic

Weaver's perspective reminds me of one seemingly off-handed remark in *Rationality: From AI to Zombies*, which jumped out at me especially for some reason. Yudkowsky notes in passing that

> I … don't know much about human anatomy with the exception of the brain. I couldn't point out on my body where my kidneys are, and I can't recall offhand what my liver does. (I am not proud of this. Alas, with all the math I need to study, I'm not likely to learn anatomy anytime soon.) (p. 167)

Now, this remark in itself isn't terribly important; and by now Yudkowsky may well have become more knowledgeable on the topic of human livers. Or the comment may even be a rhetorical flourish, with more poetic than literal truth (though I'm not sure that would be Yudkowsky's style). His remark, however, reminded me of two things:

- A paper from a couple decades ago, modeling the liver as a complex 3D self-organizing system, that recognizes toxins due to its complex self-organizing activity (Dioguardi 1989; Ramanujan 2007).

- Various perplexing reports of people, after receiving liver transplants, developing tastes similar to those of the liver donor (the liver's previous "owner") and different from their own previous tastes (Borrelli 2013).

Of course, neither of these two things is in itself particularly essential for understanding the universe, or AGI, or humanity. But they do indicate what – as I read through Yudkowsky's book – I came to see as a peculiar and systematic shortcoming in his generally polymathic perspective.

Yudkowsky isn't terribly curious about his liver or the general internal functioning of his own body. He also, as he states explicitly in his book, lacks social curiosity. One can justify this by observing that human bodies and societies are just a couple of very particular systems – so that focusing on them is like focusing on studying the chemistry of a particular kind of chocolate bar, or the highly particular flight patterns of a particular kind of insect. But then Yudkowsky also professes an outright disdain for concepts such as "complexity" and "emergence" – which he professes to see, basically, as words that people wave around when they haven't yet understood some phenomenon in fully reductionistic detail and don't want to admit how ignorant they are. And in the end, these "blind spots" appear connected to aspects of his perspective on AGI.

His view of an ideal AGI is a system that always does its best, given the computational resources available to it, to choose the actions that will maximize its utility function (which summarizes its goals) given its model of its environment. (And the AGI has built this model using internal model-building actions chosen in the same way, based on observations of the world gathered via controlling its sense-

organs with actions chosen in the same way.) Importantly, there is no room here for the AGI to encounter previously unanticipated aspects of its environment (or itself) that cause it to realize its previous goals were formulated based on a disappointingly limited understanding of the world. There is no room here for the AGI to interact with entities in the outside world in a way that causes surprising, unanticipated emergent patterns to arise, which fundamentally modify the way the AGI operates.

In Yudkowsky's idealized vision of intelligence, it seems there is no room for true *development*, in the sense in which young children develop. Development isn't just a matter of a mind learning more information or skills, or learning how to achieve its goals better. Development is a matter of a mind becoming interested in fundamentally different things. Development is triggered, in the child's mind, by a combination of what the child has become (via its own learning processes, its own goal-seeking and its own complex self-organization) and the infusion of external information. That is, development occurs when segments of the child's DNA, which have functions that manifest only once certain trigger conditions in the child's growth have been met, interact with the external world, such as other people who teach the child more based on what the child appears to know at each stage, and the physical world, which provides patterns and problems at various levels of complexity, so that as a mind advances it notices new aspects of the world suitable to provide it with interesting challenges.

An AGI may also develop in this sense – via emergent dynamics resulting from its own ongoing learning and self-organization, and information coming into it from the outside world. But development is necessarily unpredictable from the perspective of the developing system, and from the view of simpler systems in the surrounding world. An AGI that is configured to keep maximizing its original goals, no matter what, is not going to develop. It will not be an open-ended intelligence.

11. Proactionary versus precautionary principles

The notion of open-ended intelligence ties in with a very different, historically important line of thinking pursued by cross-disciplinary innovator Max More (arguably the founding philosopher of transhumanism). More has made a point of distinguishing two different principles from which future technologies and other transformational advances can be viewed (More 2005):

- **Precautionary Principle**: there is a responsibility to intervene and protect the public from exposure to harm where scientific investigation discovers a plausible risk. The protections that mitigate suspected risks can be relaxed only if further scientific findings emerge that more robustly support an alternative explanation.

- **Proactionary Principle**: "People's freedom to innovate technologically is highly valuable, even critical, to humanity. This implies several imperatives when restrictive measures are proposed: Assess risks and opportunities according to available science, not popular perception. Account for both the costs of the restrictions themselves, and those of opportunities foregone. Favor measures that are proportionate to the probability and magnitude of impacts, and that have a high expectation value. Protect people's freedom to experiment, innovate, and progress." (More 2005)

Of course, these general principles don't tell you exactly what to do in any particular case. Bostrom might argue that the potential magnitude of impact of superintelligence is so very high that even if the probability of a bad outcome is not so high, it's worth taking extreme measures. But certainly it seems clear that Bostrom's *attitude* – e.g. his choice of where to focus – is much more precautionary than proactionary.

Yudkowsky's ethos is very explicitly precautionary. In the intellectual-autobiography section of *AI to Zombies*, near the end of the massive book, he recounts the moment when he first fully understood – conceptually and viscerally – that his AGI development work had the potential to kill all humans. Before that moment, as he tells it, the potential harm of his (or others') AGI work was a sort of intellectual abstraction to him; he didn't feel it in the way he would feel the possibility of a gun he was holding to a person's head, actually killing the person. But once he fully understood the potential for existential risk in his AGI work, he knew his #1 priority had to be *to be careful*. He knew that before proceeding with building any AGI system, or even outlining any AGI design in detail, he would need to convince himself fully that he had a way to avert this potential existential risk.

The precautionary principle is fundamentally closed-ended in spirit: it is oriented toward preservation of boundaries, rather than expansion of boundaries (which is inherently risky, as it involves doing something new, which always brings on some risks).

Obviously, some degree of precaution is necessary in life, to avoid idiotically risky behaviors. But just as obviously, the development of life on Earth, and throughout the universe, has been a story of proaction on various levels: of systems going beyond their previous boundaries and exploring new structures and dynamics, even when these posed threats to what came before.

As I often put it in public lectures: Way back when, Joe Caveman may also have said to his colleagues, "Well yes, but it's been shown that the agricultural lifestyle *could* kill everybody – so we need to be on the safe side, right?" Joe Caveman would probably have been correct in his assessment of the possibility of a bad outcome. But humanity has not developed beyond apehood by being precautionary.

Yudkowsky and Bostrom are Bayesian rationalists and they both understand probability well. They know they have no rational basis to estimate the odds that AGI will prove massively destructive; otherwise they would give numerical estimates of destructive AGI-triggered events. Instead they use terms like "probably" and "almost surely" in such contexts, to indicate that what they are expressing as near-certainties are actually just their intuitions.

A more proactionary approach to superintelligence is to explicitly recognize that superintelligent AGIs will likely leave their explicit goal systems behind as they self-modify, but they may be useful scaffolding to use to get from the early stages to the later ones. One doesn't need to think about the process in terms of specific goals, but can rather think about building AGIs which are intrinsically cooperative. Intuitively, it seems that a cooperative system is open to interaction and feedback, and will likely be less prone to develop autistic or hilariously distorted interpretations of its initial goals.

As an example of the conceptual contortions that an overly precautionary perspective can lead to, consider Bostrom's example of the Paperclip Maximizer (often cited by Yudkowsky as well) – a superintelligence whose only goal is to turn everything in the universe into paperclips. A commonly discussed variation – exemplifying the concept of "perverse instantiation" discussed above – is a superintelligence whose goal is to make humans happy, and which ends up tiling the universe with little smiley-faces (because it was trained to pursue happiness via seeking to make as many smiling faces as possible).

Jack Williamson's classic SF novel *The Humanoids* (first Orb ed. 1996) explored a less ridiculous version of this situation – advanced AIs with the goal "to serve and protect and guard men from harm," who ended up drugging people with special chemicals designed to prevent them from having negative, harmful thoughts. For a long time this novel was required reading of all AI students at MIT.

I debated the Paperclip Maximizer concept at fair length with SIAI leader Luke Muehlhauser in 2012 (actually we discussed a Mickey Mouse Maximizer, but the idea was the same; Goertzel and Muehlhauser 2012). My basic argument was simple: I found it very unlikely that a superintelligent mind would orient its behavior toward such an extremely stupid goal. Given the nature of self-organization and emergence, it seems very unlikely that a complex subsystem of the universe would self-organize into a configuration in which there was a massively superhumanly intelligent system with such a stupid goal. Here the orthogonality thesis leads to a precautionary perspective in a way that feels to me particularly absurd.

Concern about the Paperclip Maximizer and its ilk appears to be a sort of pathology of extreme precautionary thinking – a case of proposing to tightly control potentially very beneficial technologies due to fear of hypothetical, theoretical worst-case outcomes whose possibility is predicated on highly dubious philosophical argumentation. As a thought experiment, the Paperclip Maximizer is interesting and stimulating. As a scare story, used as part of a campaign for tight government regulation of AGI and potential restriction of AGI R&D to a small elite group, I find it considerably less appealing.

12. The birth of the Singularity via the Emergent Global Brain?

Weaver, the cross-disciplinary thinker whose work on open-ended intelligence I've been discussing and improvising on above, is employed in a small but vibrant group called the Global Brain Institute (GBI), housed at the Free University of Brussels (VUB). One of the perspectives explored by the thinkers at the GBI is that early-stage AGI systems will interact with groups of humans and narrow AIs to form a global emergent intelligent network that can be thought of as a "global brain." This global brain, rather than any individual AGI software system or robot programmed by a particular human, will be the general intelligence that ascends to superintelligence and realizes the Singularity vision.

This possibility is worth reflecting on, in the context of the fundamentally open, transhuman nature of our apparent future. How does this kind of vision tie in with the Bostrom/Yudkowsky perspective? What would be the goals of such a global brain? What would be its utility function? Are these concepts the right ones for thinking about this sort of complex, self-organizing adaptive system – or various superintelligences it might evolve into?

Futurist AI and physics researcher Hugo de Garis likes to point out that, since human brains utilize so little of the computing power implicit in the mass-energy they are composed of, eventually any hybrid intelligent system composed of human and nonhuman components must go one of two directions (de Garis 2005):

- Stultify the potential of its nonhuman component, so as to keep it in some sort of balance with its human component; or,

- Become essentially a nonhuman system, with its human component a trivially small fraction (e.g. a cyborg whose robot brain does all but one quadrillionth of the thinking, or a Global Brain whose nonhuman computational components do all but one quadrillionth of the thinking, etc.

There is a definite point here. However, there is also an obvious counterpoint: As humans have not made ants or bacteria obsolete, superhuman AGIs need not make human beings or significantly-human-focused global brains obsolete. Furthermore, both current humans and near-future global brains or (for that matter) cyborgs may play some important role in shaping the initial launch of superintelligence.

As Ray Kurzweil has pointed out so many times, the prospect of a "machine takeover" has a whole different aspect if it's humans who are becoming the machines, at a pace slow enough that it comes with

the feeling of transitioning to a new kind of humanity, rather than the feeling of being supplanted by something alien. And since humans are not simplistic utility maximizers, there is no particular reason to suspect we will grow into that sort of pathological superintelligent system.

13. A typology of perspectives on the future of AGI

Summing up many of the themes I've explored above, I find it useful to think about the future of AGI in terms of three types of goals, and three types of agency.

The three types of goals are

1. Personal goals – for ourselves and our friends and family.

2. Humanity-wide goals – for the present and future human race.

3. Cosmist goals – for the spread of whatever values we choose around the cosmos … intelligence, life, joy, growth, discovery, etc.

The three types of agency are:

1. Personal, individual agency – what we choose to do ourselves.

2. Agency via one's pragmatic, viable influence – what we can nudge others to do via our actions.

3. Humanity-wide agency – what one would do, if one were magically made an all-powerful world dictator.

Some of the differences of opinion one sees regarding the future of AGI can be understood via noting that different people are assuming different types of goals or agency.

Three types of goals

Let's review the goals first.

Each of us has our own personal goals, which we spend a lot of our energy pursuing. For each of us, as an individual human, there is the goal of preserving our own life – of maximizing our own joy, growth and choice, or whatever other values mean a lot to us. There is also the related goal of preserving life, and positive values for our life, and the lives of our offspring specifically (rather than humanity in general). For individuals who aren't traditionally religious, and who enjoy life, maximization of healthy lifespan naturally emerges as an important personal goal.

Bostrom's *Superintelligence* considers mainly humanity-wide goals. And it's valuable to have such a careful analysis worked out from that point of view. Yet this is far from the only relevant, interesting perspective.

A broader set of goals, going beyond the scope of current human reality, is given by the Cosmist tradition (as summarized, e.g., in the Ten Cosmist Convictions that Giulio Prisco and I articulated; see Goertzel 2010). In this perspective one views general values like Life, Joy, Growth, Choice, Discovery and so forth as more important than the particulars of current humanity.

Yudkowsky has often pointed out that these "general values," as humans conceive them, are actually quite complex and not captured by simple mathematical or verbal formulations – what we mean by "Growth" actually wraps up a huge amount of human culture and psychology. This is a worthwhile point. On the other hand, the values passed along to AGIs (or cognitively enhanced humans, or human uploads, etc.) by humans don't need to be simple mathematical or verbal formulations either. The super-AGI that implacably pursues a simple mathematically formulated goal is something of a straw-man. Actual AGIs are almost sure to form their goal systems – insofar as they have explicit "goal systems" – via a combination of their initial programming and ongoing interaction with humans in human society. The complexity of human values – which are ever-changing – will morph into the complexity of post-human values. Not exactly as caveman values gradually morphed into modern human values, but via a roughly analogous process.

Absolutely, we do not know what will happen when human-level AGIs are developed. But humanity has never known what would happen when it created something dramatically new. That is the nature of dramatic progress. Personally, I would like to see humans become more and more intelligent alongside machines, and upload themselves into machines, and combine themselves with machines via cyborgization – so that the advanced AGIs that emerge are closely connected with us, not just "devices we have created." But this route also provides no guarantees.

Limits of control and varieties of agency

The grandest type of agency we can easily consider, where AGI and related advances are concerned, is the level of the human race as a whole. We can ask: Overall, given its current situation, what should humanity do? Or in other words: If I were a super-powerful puppet-master, with the ability to control human beings across the world to a huge degree, how would I have them behave just now?

This sort of thinking is an interesting philosophical and theoretical exercise, and can generate worthwhile ideals that can guide practical actions. But the dangers of this sort of thinking are also well-known. Communism arose from this kind of thinking; and if you think only about how people "should" in some sense interact, and how society "should" be organized, it's possible to convince oneself that communism in the original Marxian sense is a good idea. Yet, experience shows it tends not to work out so well in reality, at least not with current or historical levels of technology and unmodified human brains.

So for instance, when in Bostrom's *Superintelligence* I read recommendations like the one I quoted above regarding a top-secret UN/corporate-organized superintelligence project, I have to strongly question the practical relevance. I lived in the DC area for nine years and worked closely with various government agencies. There were some great people there. But it's extremely hard for me to see a global consortium of government-controlled agencies doing what Bostrom projects. And it seems quite possible to me that efforts in this direction could lead to terrible outcomes, much as did some of the attempts to practically implement Marx's utopic visions.

In reality (and setting aside issues of the philosophy of free will), anyone's agency is quite limited. Each of us is acting within a world in which other powerful forces are working, for their own diverse purposes, toward AGI and other advanced technologies. Even those humans who happen to lead great nations or great corporations are acting within this same world.

Even highly powerful human beings like Bill Gates and Barack Obama have relatively limited influence on the overall course of events. Either of these men could potentially make a significant difference to the future of AGI by, say, putting tens of billions (or even trillions, in Obama's case) of dollars into

- AGI research across the board;
- research on friendly, loving, helpful AGI specifically;
- longevity research, including AGI-powered and every other kind; or
- (on the negative side) research on military robots specifically.

But still, in spite of the relatively huge amounts of leverage they wield over human events, neither of these two individuals has the power to convince all the governments and corporations in the world to slow down progress on all advanced technologies.

The limited power of even the most powerful individuals on the planet is shown by Obama's difficulties in shutting down the Guantanamo prison, and the impressive, yet still limited, impact of the billions of dollars Gates has spent improving the level of medical care in sub-Saharan Africa. It is also shown by the complex compromise the Chinese government has had to make in allowing its citizens to travel abroad and access the Internet. Society has its own self-organizing directions of growth and development, along with its own characteristics that block progress in various directions. We can modulate it to some extent, but none of us – not even Obama, Gates, or the Chinese Communist Party – can lastingly hold back the flow of progress, at least not with high probability.

Sometimes Bostrom seems concerned with the question of what humanity should *ideally* do in its current situation. However, there are more subtleties lurking in this kind of consideration than Bostrom deigns to mention. As Weaver puts, it, regarding the issue of "human values" and their core and how they should best evolve in future,

> Besides relatively trivial issues like climate change, humanity is not very well equipped to answer this question. Achieving the highest "good" for most people at the basis of the utilitarian approach fits only very simple issues. The social construct of all humanity is far from being a unified individuated entity. It is more like an embryo with an indefinite gestation period. What such an embryo does depends on so many inner dependencies of complex interactions that no single answer can be given besides fulfilling the very basic conditions to sustain the continuation of the process and even there a consensus is not warranted. (personal communication)

These considerations are deep and fascinating, but to me, an even *more* interesting question is what should various (individual and institutional) actors in society now *actually* do, given the real-world context.

A matrix of perspectives on superintelligence

The different high-level perspectives on the future of AGI ensuing from these different varieties of goals and agency are summed up in the matrix below:

<table>
<tr><td colspan="2"></td><td colspan="3">GOALS</td></tr>
<tr><td colspan="2"></td><td>Personal</td><td>Humanity-wide, conservative</td><td>Cosmist</td></tr>
<tr><td rowspan="3">AGENCY</td><td>Personal</td><td rowspan="3">Advance toward longevity, mind uploading, brain enhancement etc. as rapidly as possible.

Advance AGI rapidly so long as the odds of achieving personally-important breakthroughs via AGI seem higher than the odds of AGI causing personal death or severe misfortune.</td><td colspan="2">Personal energy is best spent advocating pragmatically achievable goals.</td></tr>
<tr><td>Broad, pragmatic influence</td><td colspan="2">Balance of rapid and careful AGI (and other tech) development is likely best, due to the risk of others doing rapid, irresponsible or shortsighted AGI/tech development.</td></tr>
<tr><td>Humanity-wide</td><td>Slower overall tech development probably best.

Selectively slowing AGI development best only if AGI is really the greatest existential risk. (Bostrom focuses in this box)</td><td>Tech development should occur at a rational, measured pace – with the goal of understanding how to launch the next phase of intelligence in a high-quality way, and then doing it.</td></tr>
</table>

In the language used in this matrix, we may say that Bostrom and Yudkowsky, in their writings, have addressed mainly human-wide, conservative goals – rather than their own personal goals, or the goals of intelligence broadly conceived. Bostrom has focused, in his book, mainly on what humanity as a whole SHOULD do to achieve these goals.

It seems to me, however, that if one adopts a pragmatic rather than idealistic attitude to agency (i.e. looks at what can plausibly be done by real-world individuals or organizations, rather than what should ideally be done by humanity as a whole according to certain aspirations), then the conflict between conservative humanity-wide goals and more radical Cosmist goals largely disappears.

The radical proposals Bostrom tentatively ventures, with a view toward mitigating the risks of superintelligence, largely seem unrealistic or dangerous (UN control, restriction of AGI R&D to an elite group, etc.). And I suppose this is to be expected. As my father (the sociologist Dr. Ted Goertzel) said, in what I thought was the best line of the AGI-Impacts conference that Bostrom, his colleagues, and I co-organized at Oxford in 2012, "Has there ever in history been a situation where philosophers have averted a major human tragedy via their philosophizing?"

It was a rhetorical question. Of course not. I would suppose this is part of the reason Yudkowsky's crew has largely turned to mathematics these days. There *have* been situations in history where seemingly out-

there, abstract mathematics has suddenly turned out to be extremely practically valuable, much to everyone's surprise.

But if one sets aside untenable radical proposals, then what is left of Bostrom's practical recommendations is mainly an entreaty to proceed thoughtfully with AGI R&D and ongoingly consider the risks alongside the benefits (while also considering the risk that slowing down development will cause someone else to proceed faster with an ill-conceived AGI design). Such an entreaty is laudable but hardly original.

Whether one is mainly interested in human benefit narrowly construed or in the grand Cosmist future of the universe, in either case the most sensible medium-term goal for human society is to guide the advance of technology in a rational way that has reasonable odds of getting past the current phase of development without causing global annihilation or other horrible catastrophes. The presence of an uncertain, slippery, hard-to-quantify possibility of spawning human-destroying superintelligences cannot be denied, but is just one potentially troublesome aspect among many.

14. Inconclusive conclusions

Facing great uncertainty – which is what humanity has always faced – we must proceed by intuition as much as by detailed argumentation. Bostrom's arguments are clearly fueled by his own intuition that superintelligence is likely to be harmful to the things he values. On the other hand, my own intuition is that pursuing AGI now is more likely to do good than harm. I cannot prove this intuition, but nor can its opposite be proven, and we have to make pragmatic judgments now, within the actual world we live in, without benefit of proof – as has been the case throughout the history of humanity. Summing up all the various factors, my own judgment is that the best way to balance the different goals and types of agency at play is to pursue advanced AGI in a thoughtful but proactionary way.

My intuition also suggests that a broad-based, open pursuit of AGI is going to work out better than an elite group of uber-nerds locked in a secure installation and protected by a government-corporate conglomerate. It's true that the open approach brings risks with it – e.g. ISIS or some evil national dictator could fork an open-source AGI codebase and do terrible things with it. But my strong suspicion is that such forces will move much slower on the advanced technology front than teams of developers existing within the modern tech community. The benefit from having a large part of the global brain bearing down on the problem seems to me greater than the risk of these marginal forces summoning the brainpower and focus to make a massively-destructive AGI or proto-AGI before the international scientific/engineering/hacking community cracks the AGI problem.

Humanity emerged from apehood via proactionary manifestations of open-ended intelligence, and it seems that, overall, the later stages of progress toward Technological Singularity are currently unfolding in a similar manner. Folks like Bostrom and Yudkowsky are part of the process, and probably play a valuable role in nudging the world's attention a bit toward the risks of advanced technology, so that a little bit more care will be taken as the inexorable developments unfold. But it seems extremely unlikely that the threshold to superintelligence will be passed via Yudkowsky – or any other individual or small group – locked in a basement protected, controlled, and funded by the UN, in cahoots with national governments and major corporations.

The philosophizing of Bostrom, Yudkowsky and kin has certainly had some practical impact, in terms of guiding the attention and effort of others. The $10M allocated to "safe AI research" by Elon Musk as a fairly direct result of Bostrom's book is just the most direct example. More broadly there is no doubt that the writing and speaking of both of these men has made a difference in how the world thinks about AGI

and its potentials. What real practical difference their work will make is hard to foresee, and may be difficult to unravel even in hindsight.

But still – while Bostrom has grabbed the spotlight lately, I personally think Weaver is much more on target with his views about open-ended intelligence and the Global Brain. AGI is emerging, little by little, via the loosely-coordinated, self-organizing efforts of a massive network of people and institutions around the world. AGI is being thought-out by the Global Brain, even more so than by any individual or small group – and it's probably the most interesting thought on the mind of the Global Brain right now. The Global Brain, and the humans who play parts in it, will be massively transformed via the transition of AGI from thought into reality. There are no guarantees, there never have been any guarantees; all we can do is do our best to nudge these grand developments in directions that are positive according to our various mixes of values.

Acknowledgments: Thanks are due to Sander Olson for gifting me a copy of *Superintelligence* and encouraging me to write this review; and to Sander and Zar Goertzel for face-to-face discussions (in College Park in July 2015) that formed the core of this essay. Also to Weaver for his talk on Open-Ended Intelligence at the *Artificial General Intelligence 2015* conference, and to Weaver and Viktoras Veitas for various related face-to-face and email conversations, plus many detailed suggestions and edits to an earlier draft of this essay. And to all the other people who have indulged my passion for discussing these topics over the years.

And finally, thanks to an anonymous referee for many thoughtful suggestions which led me to add various sentences and paragraphs here and there, improving the article considerably. This was a rare occasion where refereeing was actually useful rather than just a pain or a hurdle to get through. The flaws remaining are of course my responsibility alone.

References

Allen, P.G. and M. Greaves. 2011. Paul Allen: The Singularity isn't near. *MIT Technology Review* Oct 12.

Baum, S., B. Goertzel, and T.G. Goertzel. 2011. How long until human-level AI? Results from an expert assessment. *Technological Forecasting & Social Change* 78(1):185–95.

Borrelli, L. 2013. Can an organ transplant change a recipient's personality? Cell memory theory affirms "Yes." *Medical Daily* July 9.

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

Cellan-Jones, R. 2014. Stephen Hawking warns artificial intelligence could end mankind. *BBC News* Dec 2.

Chalmers, D. 2010. The Singularity: A philosophical analysis. *Journal of Consciousness Studies* 17:7–65

de Garis, H. 2005. *The artilect war*. Etc. Publications.

Dioguardi, N. 1989. The liver as a self-organizing system. I. Theoretics of its representation. *La Ricerca in Clinica e in Laboratorio* 19(4):281–99.

Goertzel, B. 2010. The Singularity Institute's scary idea. http://multiverseaccordingtoben.blogspot.hk/2010/10/singularity-institutes-scary-idea-and.html

Goertzel, B. 2010. *A cosmist manifesto*. Humanity+ Press.

Goertzel, B. 2011. Beyond nanotech to femtotech. *H+ Magazine* Jan 10. http://hplusmagazine.com/2011/01/10/theres-plenty-more-room-bottom-beyond-nanotech-femtotech/

Goertzel, B. 2012. Should humanity build a global AI nanny to delay the Singularity until it's better understood? *Journal of Consciousness Studies* 19(1 –2): 96–111

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

Goertzel. B. and J. Pitt. 2012. Nine ways to bias open-source AGI toward Friendliness. *Journal of Evolution and Technology* 22(1):116–131.

Holley, 2015. Bill Gates on dangers of artificial intelligence: "I don't understand why some people are not concerned." *Wash. Post* Jan. 29.

Joy, B. 2001. Why the future doesn't need us. *Wired Magazine* April.

Kumparak, G. 2014. Elon Musk compares building artificial intelligence to "Summoning the demon." *TechCrunch* Oct 26.

Lawrence, J. 2013. 3 Years on, questions remain on Thiel Fellowship and entrepreneurship. *Education News* Sep 16.

More, M, 2005. The Proactionary Principle. http://www.maxmore.com/proactionary.html.

Omohundro, S. 2008. The basic AI drives. In *Proceedings of the First AGI Conference, Volume 171, Frontiers in Artificial Intelligence and Applications*, edited by P. Wang, B. Goertzel, and S. Franklin, February 2008, IOS Press.

Ramanujan, V.K. and B. Herman. 2007. Aging process modulates nonlinear dynamics in liver cell metabolism. *Journal of Biological Chemistry* 282: 19217–19226.

Reinhart, B. 2011. SIAI – An examination. http://lesswrong.com/lw/5il/siai_an_examination/.

Tsegaye, H. 2015. What does the Singularity mean for Africa? In *The End of the Beginning*, ed. B. Goertzel and T. Goertzel, Humanity+ Press.

Weinbaum, D. and V. Veitas, 2015. Open ended intelligence: The individuation of intelligent agents. http://arxiv.org/abs/1505.06366.

Williamson, J. 1996. *The humanoids*. New York: Orb Books. (Based on material orig. pub. 1947–49.)

Yudkowsky, E. 2001. Staring into the Singularity. http://www.yudkowsky.net/obsolete/singularity.html.

Yudkowsky, E. 2015. *Rationality: From AI to zombies*. Machine Intelligence Research Institute.

Yudkowsky, E. 2015. Harry Potter and the methods of rationality. http://hpmor.com/.