Artificial Intelligence and Its Implications for Future Suffering

Brian Tomasik
Center on Long-Term Risk
brian.tomasik@longtermrisk.org

June 2016*

Abstract

Artificial intelligence (AI) will transform the world later this century. I expect this transition will be a "soft takeoff" in which many sectors of society update together in response to incremental AI developments, though the possibility of a harder takeoff in which a single AI project "goes foom" shouldn't be ruled out. If a rogue AI gained control of Earth, it would proceed to accomplish its goals by colonizing the galaxy and undertaking some very interesting achievements in science and engineering. On the other hand, it would not necessarily respect human values, including the value of preventing the suffering of less powerful creatures. Whether a rogue-AI scenario would entail more expected suffering than other scenarios is a question to explore further. Regardless, the field of AI ethics and policy seems to be a very important space where altruists can make a positive-sum impact along many dimensions. Expanding dialogue and challenging us-vs.-them prejudices could be valuable.

Contents

1 Summary **3**

2 Introduction **3**

3 Is "the singularity" crazy? **4**

4 The singularity is more than AI **5**

5 Will society realize the importance of AI? **6**

6 A soft takeoff seems more likely? **6**

7 Intelligence explosion? **10**

8 Reply to Bostrom's arguments for a hard takeoff **12**

*First written: May 2014; last modified: Jun. 2016

9 How complex is the brain? **15**
9.1 One basic algorithm? 15
9.2 Ontogenetic development 16

10 Brain quantity vs. quality **17**

11 More impact in hard-takeoff scenarios? **17**

12 Village idiot vs. Einstein **19**

13 A case for epistemic modesty on AI timelines **20**

14 Intelligent robots in your backyard **20**

15 Is automation "for free"? **21**

16 Caring about the AI's goals **22**

17 Rogue AI would not share our values **23**

18 Would a human-inspired AI or rogue AI cause more suffering? **24**

19 Would helper robots feel pain? **26**

20 How accurate would simulations be? **27**

21 Rogue AIs can take off slowly **28**

22 Would superintelligences become existentialists? **29**

23 AI epistemology **30**

24 Artificial philosophers **31**

25 Would all AIs colonize space? **31**

26 Who will first develop human-level AI? **33**

27 One hypothetical AI takeoff scenario **33**

28 How do you socialize an AI? **36**
28.1 Treacherous turn 39
28.2 Following role models? 40

29 AI superpowers? **40**

30 How big would a superintelligence be? **41**

31 Another hypothetical AI takeoff scenario **41**

32 AI: More like the economy than like robots? 42

33 Importance of whole-brain emulation 44

34 Why work against brain-emulation risks appeals to suffering reducers 45

35 Would emulation work accelerate neuromorphic AI? 45

36 Are neuromorphic or mathematical AIs more controllable? 46

37 Impacts of empathy for AIs 47
37.1 Slower AGI development? 47
37.2 Attitudes toward AGI control 48

38 Charities working on this issue 49

39 Is MIRI's work too theoretical? 49

40 Next steps 51

41 Where to push for maximal impact? 52

42 Is it valuable to work at or influence an AGI company? 55

43 Should suffering reducers focus on AGI safety? 56

44 Acknowledgments 57

References 57

1 Summary

Artificial intelligence (AI) will transform the world later this century. I expect this transition will be a "soft takeoff" in which many sectors of society update together in response to incremental AI developments, though the possibility of a harder takeoff in which a single AI project "goes foom" shouldn't be ruled out. If a rogue AI gained control of Earth, it would proceed to accomplish its goals by colonizing the galaxy and undertaking some very interesting achievements in science and engineering. On the other hand, it would not necessarily respect human values, including the value of preventing the suffering of less powerful creatures. Whether a rogue-AI scenario would entail more expected suffering than other scenarios is a question to explore further. Regardless, the field of AI ethics and policy seems to be a very important space where altruists can make a positive-sum impact along many dimensions. Expanding dialogue and challenging us-vs.-them prejudices could be valuable.

2 Introduction

This piece contains some observations on what looks to be potentially a coming machine revolution in Earth's history. For general back-

ground reading, a good place to start is Wikipedia's article on the technological singularity.

I am not an expert on all the arguments in this field, and my views remain very open to change with new information. In the face of epistemic disagreements with other very smart observers, it makes sense to grant some credence to a variety of viewpoints. Each person brings unique contributions to the discussion by virtue of his or her particular background, experience, and intuitions.

To date, I have not found a detailed analysis of how those who are moved more by preventing suffering than by other values should approach singularity issues. This seems to me a serious gap, and research on this topic deserves high priority. In general, it's important to expand discussion of singularity issues to encompass a broader range of participants than the engineers, technophiles, and science-fiction nerds who have historically pioneered the field.

I. J. Good (1982) observed: "The urgent drives out the important, so there is not very much written about ethical machines". Fortunately, this may be changing.

3 Is "the singularity" crazy?

In fall 2005, a friend pointed me to Ray Kurzweil's (2000) *The Age of Spiritual Machines*. This was my first introduction to "singularity" ideas, and I found the book pretty astonishing. At the same time, much of it seemed rather implausible to me. In line with the attitudes of my peers, I assumed that Kurzweil was crazy and that while his ideas deserved further inspection, they should not be taken at face value.

In 2006 I discovered Nick Bostrom and Eliezer Yudkowsky, and I began to follow the organization then called the Singularity Institute for Artificial Intelligence (SIAI), which is now MIRI. I took SIAI's ideas more seriously than Kurzweil's, but I remained embarrassed to mention the organization because the first word in SIAI's name sets off "insanity alarms" in listeners.

I began to study machine learning in order to get a better grasp of the AI field, and in fall 2007, I switched my college major to computer science. As I read textbooks and papers about machine learning, I felt as though "narrow AI" was very different from the strong-AI fantasies that people painted. "AI programs are just a bunch of hacks," I thought. "This isn't intelligence; it's just people using computers to manipulate data and perform optimization, and they dress it up as 'AI' to make it sound sexy." Machine learning in particular seemed to be just a computer scientist's version of statistics. Neural networks were just an elaborated form of logistic regression. There were stylistic differences, such as computer science's focus on cross-validation and bootstrapping instead of testing parametric models – made possible because computers can run data-intensive operations that were inaccessible to statisticians in the 1800s. But overall, this work didn't seem like the kind of "real" intelligence that people talked about for general AI.

This attitude began to change as I learned more cognitive science. Before 2008, my ideas about human cognition were vague. Like most science-literate people, I believed the brain was a product of physical processes, including firing patterns of neurons. But I lacked further insight into what the black box of brains might contain. This led me to be confused about what "free will" meant until mid-2008 and about what "consciousness" meant until late 2009. Cognitive science showed me that the brain was in fact very much like a computer, at least in the sense of being a deterministic information-processing device with distinct algorithms and modules. When viewed up close, these algorithms could look as "dumb" as the kinds of algorithms in narrow AI that I had previously dismissed as "not really intelligence." Of course, animal brains combine

these seemingly dumb subcomponents in dazzlingly complex and robust ways, but I could now see that the difference between narrow AI and brains was a matter of degree rather than kind. It now seemed plausible that broad AI could emerge from lots of work on narrow AI combined with stitching the parts together in the right ways.

So the singularity idea of artificial general intelligence seemed less crazy than it had initially. This was one of the rare cases where a bold claim turned out to look *more* probable on further examination; usually extraordinary claims lack much evidence and crumble on closer inspection. I now think it's quite likely (maybe $\sim 75\%$) that humans will produce at least a human-level AI within the next ~ 300 years conditional on no major disasters (such as sustained world economic collapse, global nuclear war, large-scale nanotech war, etc.), and also ignoring anthropic considerations (Bostrom, 2010).

4 The singularity is more than AI

The "singularity" concept is broader than the prediction of strong AI and can refer to several distinct sub-meanings. Like with most ideas, there's a lot of fantasy and exaggeration associated with "the singularity," but at least the core idea that technology will progress at an accelerating rate for some time to come, absent major setbacks, is not particularly controversial. Exponential growth is the standard model in economics, and while this can't continue forever, it has been a robust pattern throughout human and even pre-human history.

MIRI emphasizes AI for a good reason: At the end of the day, the long-term future of our galaxy will be dictated by AI, not by biotech, nanotech, or other lower-level systems. AI is the "brains of the operation." Of course, this doesn't automatically imply that AI should be the primary focus of our attention. Maybe other revolutionary technologies or social forces will come first and deserve higher priority. In practice, I think focusing on AI specifically seems quite important even relative to competing scenarios, but it's good to explore many areas in parallel to at least a shallow depth.

In addition, I don't see a sharp distinction between "AI" and other fields. Progress in AI software relies heavily on computer hardware, and it depends at least a little bit on other fundamentals of computer science, like programming languages, operating systems, distributed systems, and networks. AI also shares significant overlap with neuroscience; this is especially true if whole brain emulation arrives before bottom-up AI. And everything else in society matters a lot too: How intelligent and engineering-oriented are citizens? How much do governments fund AI and cognitive-science research? (I'd encourage less rather than more.) What kinds of military and commercial applications are being developed? Are other industrial backbone components of society stable? What memetic lenses does society have for understanding and grappling with these trends? And so on. The AI story is part of a larger story of social and technological change, in which one part influences other parts.

Significant trends in AI may not look like the AI we see in movies. They may not involve animal-like cognitive agents as much as more "boring", business-oriented computing systems. Some of the most transformative computer technologies in the period 2000-2014 have been drones, smart phones, and social networking. These all involve some AI, but the AI is mostly used as a component of a larger, non-AI system, in which many other facets of software engineering play at least as much of a role.

Nonetheless, it seems nearly inevitable to me that digital intelligence in some form will eventually leave biological humans in the dust, *if* technological progress continues without fal-

tering. This is almost obvious when we zoom out and notice that the history of life on Earth consists in one species outcompeting another, over and over again. Ecology's competitive exclusion principle suggests that in the long run, either humans or machines will ultimately occupy the role of the most intelligent beings on the planet, since "When one species has even the slightest advantage or edge over another then the one with the advantage will dominate in the long term."

5 Will society realize the importance of AI?

The basic premise of superintelligent machines who have different priorities than their creators has been in public consciousness for many decades. Arguably even *Frankenstein*, published in 1818, expresses this basic idea, though more modern forms include *2001: A Space Odyssey* (1968), *The Terminator* (1984), *I, Robot* (2004), and many more. Probably most people in Western countries have at least heard of these ideas if not watched or read pieces of fiction on the topic.

So why do most people, including many of society's elites, ignore strong AI as a serious issue? One reason is just that the world is really big, and there are many important (and not-so-important) issues that demand attention. Many people think strong AI is too far off, and we should focus on nearer-term problems. In addition, it's possible that science fiction itself is part of the reason: People may write off AI scenarios as "just science fiction," as I would have done prior to late 2005. (Of course, this is partly for good reason, since depictions of AI in movies are usually very unrealistic.) Often, citing Hollywood is taken as a thought-stopping deflection of the possibility of AI getting out of control, without much in the way of substantive argument to back up that stance. For example: "let's please keep the discussion firmly within the realm of reason and leave the robot uprisings to Hollywood screenwriters."

As AI progresses, I find it hard to imagine that mainstream society will ignore the topic forever. Perhaps awareness will accrue gradually, or perhaps an AI Sputnik moment will trigger an avalanche of interest. Stuart Russell expects that

> Just as nuclear fusion researchers consider the problem of *containment* of fusion reactions as one of the primary problems of their field, it seems inevitable that issues of control and safety will become central to AI as the field matures.

I think it's likely that issues of AI policy will be debated heavily in the coming decades, although it's possible that AI will be like nuclear weapons – something that everyone is afraid of but that countries can't stop because of arms-race dynamics. So even if AI proceeds slowly, there's probably value in thinking more about these issues well ahead of time, though I wouldn't consider the counterfactual value of doing so to be astronomical compared with other projects in part because society will pick up the slack as the topic becomes more prominent.

[*Update, Feb. 2015*: I wrote the preceding paragraphs mostly in May 2014, before Nick Bostrom's *Superintelligence* book was released. Following Bostrom's book, a wave of discussion about AI risk emerged from Elon Musk, Stephen Hawking, Bill Gates, and many others. AI risk suddenly became a mainstream topic discussed by almost every major news outlet, at least with one or two articles. This foreshadows what we'll see more of in the future. The outpouring of publicity for the AI topic happened far sooner than I imagined it would.]

6 A soft takeoff seems more likely?

Various thinkers have debated the likelihood of a "hard" takeoff – in which a single com-

puter or set of computers rapidly becomes superintelligent on its own – compared with a "soft" takeoff – in which society as a whole is transformed by AI in a more distributed, continuous fashion. "The Hanson-Yudkowsky AI-Foom Debate" discusses this in great detail (Hanson & Yudkowsky, 2013). The topic has also been considered by many others, such as Ramez Naam vs. William Hertling.

For a long time I inclined toward Yudkowsky's vision of AI, because I respect his opinions and didn't ponder the details too closely. This is also the more prototypical example of rebellious AI in science fiction. In early 2014, a friend of mine challenged this view, noting that computing power is a severe limitation for human-level minds. My friend suggested that AI advances would be slow and would diffuse through society rather than remaining in the hands of a single developer team. As I've read more AI literature, I think this soft-takeoff view is pretty likely to be correct. Science is always a gradual process, and almost all AI innovations historically have moved in tiny steps. I would guess that even the evolution of humans from their primate ancestors was a "soft" takeoff in the sense that no single son or daughter was vastly more intelligent than his or her parents. The evolution of technology in general has been fairly continuous. I probably agree with Paul Christiano that "it is unlikely that there will be rapid, discontinuous, and unanticipated developments in AI that catapult it to superhuman levels [...]."

Of course, it's not guaranteed that AI innovations will diffuse throughout society. At some point perhaps governments will take control, in the style of the Manhattan Project, and they'll keep the advances secret. But even then, I expect that the internal advances by the research teams will add cognitive abilities in small steps. Even if you have a theoretically optimal intelligence algorithm, it's constrained by computing resources, so you either need lots of hardware or approximation hacks (or most likely both) before it can function effectively in the high-dimensional state space of the real world, and this again implies a slower trajectory. Marcus Hutter's AIXI(tl) is an example of a theoretically optimal general intelligence, but most AI researchers feel it won't work for artificial general intelligence (AGI) because it's astronomically expensive to compute. Ben Goertzel explains: "I think that tells you something interesting. It tells you that dealing with resource restrictions – with the boundedness of time and space resources – is actually critical to intelligence. If you lift the restriction to do things efficiently, then AI and AGI are trivial problems."[1]

In "I Still Don't Get Foom", Robin Hanson contends:

> Yes, sometimes architectural choices have wider impacts. But I was an artificial intelligence researcher for nine years, ending twenty years ago, and I never saw an architecture choice make a huge difference, relative to other reasonable architecture choices. For most big systems, overall architecture matters a lot less than getting lots of detail right.

This suggests that it's unlikely that a single insight will make an astronomical difference to an AI's performance.

Similarly, my experience is that machine-learning algorithms matter less than the data they're trained on. I think this is a general sentiment among data scientists. There's a famous slogan that "More data is better data." A main reason Google's performance is so

[1] Stuart Armstrong agrees that AIXI probably isn't a feasible approach to AGI, but he feels there might exist other, currently undiscovered mathematical insights like AIXI that could yield AGI in a very short time span. Maybe, though I think this is pretty unlikely. I suppose at least a few people should explore these scenarios, but plausibly most of the work should go toward pushing on the more likely outcomes.

good is that it has so many users that even obscure searches, spelling mistakes, etc. will appear somewhere in its logs. But if many performance gains come from data, then they're constrained by hardware, which generally grows steadily.

Hanson's "I Still Don't Get Foom" post continues: "To be much better at learning, the project would instead have to be much better at hundreds of specific kinds of learning. Which is very hard to do in a small project." Anders Sandberg makes a similar point:

> As the amount of knowledge grows, it becomes harder and harder to keep up and to get an overview, necessitating specialization. [...] This means that a development project might need specialists in many areas, which in turns means that there is a lower size of a group able to do the development. In turn, this means that it is very hard for a small group to get far ahead of everybody else in all areas, simply because it will not have the necessary know how in all necessary areas. The solution is of course to hire it, but that will enlarge the group.

One of the more convincing anti-"foom" arguments is J. Storrs Hall's (2008) point that an AI improving itself to a world superpower would need to outpace *the entire world economy* of 7 billion people, plus natural resources and physical capital. It would do much better to specialize, sell its services on the market, and acquire power/wealth in the ways that most people do. There are plenty of power-hungry people in the world, but usually they go to Wall Street, K Street, or Silicon Valley rather than trying to build world-domination plans in their basement. Why would an AI be different? Some possibilities:

1. By being built differently, it's able to concoct an effective world-domination strategy that no human has thought of.
2. Its non-human form allows it to diffuse throughout the Internet and make copies of itself.

I'm skeptical of #1, though I suppose if the AI is very alien, these kinds of unknown unknowns become more plausible. #2 is an interesting point. It seems like a pretty good way to spread yourself as an AI is to become a useful software product that lots of people want to install, i.e., to sell your services on the world market, as Hall said. Of course, once that's done, perhaps the AI could find a way to take over the world. Maybe it could silently quash competitor AI projects. Maybe it could hack into computers worldwide via the Internet and Internet of Things, as the AI did in the *Delete* series. Maybe it could devise a way to convince humans to give it access to sensitive control systems, as Skynet did in *Terminator 3*.

I find these kinds of scenarios for AI takeover more plausible than a rapidly self-improving superintelligence. Indeed, even a human-level intelligence that can distribute copies of itself over the Internet might be able to take control of human infrastructure and hence take over the world. No "foom" is required.

Rather than discussing hard-vs.-soft takeoff arguments more here, I added discussion to Wikipedia where the content will receive greater readership. See "Hard vs. soft takeoff" in "Recursive self-improvement".

The hard vs. soft distinction is obviously a matter of degree. And maybe *how long* the process takes isn't the most relevant way to slice the space of scenarios. For practical purposes, the more relevant question is: Should we expect control of AI outcomes to reside primarily in the hands of a few "seed AI" developers? In this case, altruists should focus on influencing a core group of AI experts, or maybe their military/corporate leaders. Or should we expect that society as a whole will play a big role in shaping how AI is developed and used? In this case, governance structures, social dy-

namics, and non-technical thinkers will play an important role not just in influencing how much AI research happens but also in how the technologies are deployed and incrementally shaped as they mature.

It's possible that one country – perhaps the United States, or maybe China in later decades – will lead the way in AI development, especially if the research becomes nationalized when AI technology grows more powerful. Would this country then take over the world? I'm not sure. The United States had a monopoly on nuclear weapons for several years after 1945, but it didn't bomb the Soviet Union out of existence. A country with a monopoly on artificial superintelligence might refrain from destroying its competitors as well. On the other hand, AI should enable vastly more sophisticated surveillance and control than was possible in the 1940s, so a monopoly might be sustainable even without resorting to drastic measures. In any case, perhaps a country with superintelligence would just economically outcompete the rest of the world, rendering military power superfluous.

Besides a single country taking over the world, the other possibility (perhaps more likely) is that AI is developed in a distributed fashion, either openly as is the case in academia today, or in secret by governments as is the case with other weapons of mass destruction.

Even in a soft-takeoff case, there would come a point at which humans would be unable to keep up with the pace of AI thinking. (We already see an instance of this with algorithmic stock-trading systems, although human traders are still needed for more complex tasks right now.) The reins of power would have to be transitioned to faster human uploads, trusted AIs built from scratch, or some combination of the two. In a slow scenario, there might be many intelligent systems at comparable levels of performance, maintaining a balance of power, at least for a while.[2] In the long run, a singleton (Bostrom, 2006) seems plausible because computers – unlike human kings – can reprogram their servants to want to obey their bidding, which means that as an agent gains more central authority, it's not likely to later lose it by internal rebellion (only by external aggression).

Most of humanity's problems are fundamentally coordination problems / selfishness problems. If humans were perfectly altruistic, we could easily eliminate poverty, overpopulation, war, arms races, and other social ills. There would remain "man vs. nature" problems, but these are increasingly disappearing as technology advances. Assuming a digital singleton emerges, the chances of it going extinct seem very small (except due to alien invasions or other external factors) because unless the singleton has a very myopic utility function, it should consider carefully all the consequences of its actions – in contrast to the "fools rush in" approach that humanity currently takes toward most technological risks, due to wanting the benefits of and profits from technology right away and not wanting to lose out to competitors. For this reason, I suspect that most of George Dvorsky's "12 Ways Humanity Could Destroy The Entire Solar System" are unlikely to happen, since most of them presuppose blundering by an advanced Earth-originating intelligence, but probably by the time Earth-originating intelligence would be able to carry out interplanetary engineering on a nontrivial scale, we'll already have a digital singleton that thoroughly explores the risks of

[2]Marcus Hutter imagines a society of AIs that compete for computing resources in a similar way as animals compete for food and space. Or like corporations compete for employees and market share. He suggests that such competition might render initial conditions irrelevant. Maybe, but it's also quite plausible that initial conditions would matter a lot. Many evolutionary pathways depended sensitively on particular events – e.g., asteroid impacts – and the same is true for national, corporate, and memetic power.

its actions before executing them. That said, this might not be true if competing AIs begin astroengineering before a singleton is completely formed. (By the way, I should point out that I prefer it if the cosmos isn't successfully colonized, because doing so is likely to astronomically multiply sentience and therefore suffering.)

7 Intelligence explosion?

Sometimes it's claimed that we should expect a hard takeoff because AI-development dynamics will fundamentally change once AIs can start improving themselves. One stylized way to explain this is via differential equations. Let I(t) be the intelligence of AIs at time t.

- While humans are building AIs, we have, $dI/dt=c$, where c is some constant level of human engineering ability. This implies $I(t)=ct+\text{constant}$, a linear growth of I with time.
- In contrast, once AIs can design themselves, we'll have $dI/dt=kI$ for some k. That is, the rate of growth will be faster as the AI designers become more intelligent. This implies $I(t)=Ae^t$ for some constant A.

Luke Muehlhauser reports that the idea of intelligence explosion once machines can start improving themselves "ran me over like a train. Not because it was absurd, but because it was clearly true." I think this kind of exponential feedback loop is the basis behind many of the intelligence-explosion arguments.

But let's think about this more carefully. What's so special about the point where machines can understand and modify themselves? Certainly understanding your own source code helps you improve yourself. But humans *already* understand the source code of present-day AIs with an eye toward improving *it*. Moreover, present-day AIs are vastly simpler than human-level ones will be, and present-day AIs are far less intelligent than the humans who create them. Which is easier: (1) improving the intelligence of something as smart as you, or (2) improving the intelligence of something far dumber? (2) is usually easier. So if anything, AI intelligence should be "exploding" faster now, because it can be lifted up by something vastly smarter than it. Once AIs need to improve themselves, they'll have to pull up on their own bootstraps, without the guidance of an already existing model of far superior intelligence on which to base their designs.

As an analogy, it's harder to produce novel developments if you're the market-leading company; it's easier if you're a competitor trying to catch up, because you know what to aim for and what kinds of designs to reverse-engineer. AI right now is like a competitor trying to catch up to the market leader.

Another way to say this: The constants in the differential equations might be important. Even if human AI-development progress is linear, that progress might be faster than a slow exponential curve until some point far later where the exponential catches up.

In any case, I'm cautious of simple differential equations like these. Why should the rate of intelligence increase be proportional to the intelligence level? Maybe the problems become much harder at some point. Maybe the systems become fiendishly complicated, such that even small improvements take a long time. Robin Hanson echoes this suggestion:

> Students get smarter as they learn more, and learn how to learn. However, we teach the most valuable concepts first, and the productivity value of schooling eventually falls off, instead of exploding to infinity. Similarly, the productivity improvement of factory workers typically slows with time, following a power law.
>
> At the world level, average IQ scores have increased dramatically over the last century (the Flynn effect), as the world

> has learned better ways to think and to teach. Nevertheless, IQs have improved steadily, instead of accelerating. Similarly, for decades computer and communication aids have made engineers much "smarter," without accelerating Moore's law. While engineers got smarter, their design tasks got harder.

Also, ask yourself this question: Why do startups exist? Part of the answer is that they can innovate faster than big companies due to having less institutional baggage and legacy software.[3] It's harder to make radical changes to big systems than small systems. Of course, like the economy does, a self-improving AI could create its own virtual startups to experiment with more radical changes, but just as in the economy, it might take a while to prove new concepts and then transition old systems to the new and better models.

In discussions of intelligence explosion, it's common to approximate AI productivity as scaling linearly with number of machines, but this may or may not be true depending on the degree of parallelizability. Empirical examples for human-engineered projects show diminishing returns with more workers, and while computers may be better able to partition work due to greater uniformity and speed of communication, there will remain some overhead in parallelization. Some tasks may be inherently non-paralellizable, preventing the kinds of ever-faster performance that the most extreme explosion scenarios envisage.

Fred Brooks's (1995) "No Silver Bullet" paper argued that "there is no single development, in either technology or management technique, which by itself promises even one order of magnitude improvement within a decade in productivity, in reliability, in simplicity." Likewise, Wirth's law reminds us of how fast software complexity can grow. These points make it seem less plausible that an AI system could rapidly bootstrap itself to superintelligence using just a few key as-yet-undiscovered insights.

Eventually there has to be a leveling off of intelligence increase if only due to physical limits. On the other hand, one argument in favor of differential equations is that the economy has fairly consistently followed exponential trends since humans evolved, though the exponential growth rate of today's economy remains small relative to what we typically imagine from an "intelligence explosion".

I think a stronger case for intelligence explosion is the clock-speed difference between biological and digital minds (Sotala, 2012). Even if AI development becomes very slow in subjective years, once AIs take it over, in objective years (i.e., revolutions around the sun), the pace will continue to look blazingly fast. But if enough of society is digital by that point (including human-inspired subroutines and maybe full digital humans), then digital speedup won't give a unique advantage to a single AI project that can then take over the world. Hence, hard takeoff in the sci fi sense still isn't guaranteed. Also, Hanson argues that faster minds would produce a one-time jump in economic output but not necessarily a sustained higher *rate* of growth.

Another case for intelligence explosion is that intelligence growth might not be driven by the intelligence of a given agent so much as by the collective man-hours (or machine-hours) that would become possible with more resources. I suspect that AI research could accelerate at least 10 times if it had 10-50 times more funding. (This is not the same as saying I want funding increased; in fact, I probably want funding decreased to give society more time to sort through these issues.) The population of digital minds that could be created in a few decades might exceed the biological

[3] Another part of the answer has to do with incentive structures – e.g., a founder has more incentive to make a company succeed if she's mainly paid in equity than if she's paid large salaries along the way.

human population, which would imply faster progress if only by numerosity. Also, the digital minds might not need to sleep, would focus intently on their assigned tasks, etc. However, once again, these are advantages in objective time rather than collective subjective time. And these advantages would not be uniquely available to a single first-mover AI project; any wealthy and technologically sophisticated group that wasn't too far behind the cutting edge could amplify its AI development in this way.

(A few weeks after writing this section, I learned that Ch. 4 of Nick Bostrom's (2014) *Superintelligence: Paths, Dangers, Strategies* contains surprisingly similar content, even up to the use of dI/dt as the symbols in a differential equation. However, Bostrom comes down mostly in favor of the likelihood of an intelligence explosion. I reply to Bostrom's arguments in the next section.)

8 Reply to Bostrom's arguments for a hard takeoff

In Ch. 4 of *Superintelligence*, Bostrom suggests several factors that might lead to a hard or at least semi-hard takeoff. I don't fully disagree with his points, and because these are difficult issues, I agree that Bostrom might be right. But I want to play devil's advocate and defend the soft-takeoff view. I've distilled and paraphrased what I think are 6 core arguments, and I reply to each in turn.

#1: There might be a key missing algorithmic insight that allows for dramatic progress.

Maybe, but do we have much precedent for this? As far as I'm aware, all individual AI advances – and indeed, most technology advances in general – have not represented astronomical improvements over previous designs. Maybe connectionist AI systems represented a game-changing improvement *relative to* symbolic AI for messy tasks like vision, but I'm not sure how much of an improvement they represented relative to the best alternative technologies. After all, neural networks are in some sense just fancier forms of pre-existing statistical methods like logistic regression. And even neural networks came in stages, with the perceptron, multi-layer networks, backpropagation, recurrent networks, deep networks, etc. The most groundbreaking machine-learning advances may reduce error rates by a half or something, which may be commercially very important, but this is not many orders of magnitude as hard-takeoff scenarios tend to assume.

Outside of AI, the Internet changed the world, but it was an accumulation of many insights. Facebook has had massive impact, but it too was built from many small parts and grew in importance slowly as its size increased. Microsoft became a virtual monopoly in the 1990s but perhaps more for business than technology reasons, and its power in the software industry at large is probably not growing. Google has a quasi-monopoly on web search, kicked off by the success of PageRank, but most of its improvements have been small and gradual. Google has grown very powerful, but it hasn't maintained a permanent advantage that would allow it to take over the software industry.

Acquiring nuclear weapons might be the closest example of a single discrete step that most dramatically changes a country's position, but this may be an outlier. Maybe other advances in weaponry (arrows, guns, etc.) historically have had somewhat dramatic effects.

Bostrom doesn't present specific arguments for thinking that a few crucial insights may produce radical jumps. He suggests that we might not notice a system's improvements until it passes a threshold, but this seems absurd, because at least the AI developers would need to be intimately acquainted with the AI's performance. While not strictly accurate, there's a slogan: "You can't improve what you can't measure." Maybe the AI's

progress wouldn't make world headlines, but the academic/industrial community would be well aware of nontrivial breakthroughs, and the AI developers would live and breathe performance numbers.

#2: Once an AI passes a threshold, it might be able to absorb vastly more content (e.g., by reading the Internet) that was previously inaccessible.

Absent other concurrent improvements I'm doubtful this would produce take-over-the-world superintelligence, because the world's current superintelligence (namely, humanity as a whole) already has read most of the Internet – indeed, has written it. I guess humans haven't read automatically generated text/numerical context, but the insights gleaned purely from reading such material would be low without doing more sophisticated data mining and learning on top of it, and presumably such data mining would have already been in progress well before Bostrom's hypothetical AI learned how to read.

In any case, I doubt reading with understanding is such an all-or-nothing activity that it can suddenly "turn on" once the AI achieves a certain ability level. As Bostrom says (p. 71), reading with the comprehension of a 10-year-old is probably AI-complete, i.e., requires solving the general AI problem. So assuming that you can switch on reading ability with one improvement is equivalent to assuming that a single insight can produce astronomical gains in AI performance, which we discussed above. If that's not true, and if before the AI system with 10-year-old reading ability was an AI system with a 6-year-old reading ability, why wouldn't that AI have already devoured the Internet? And before that, why wouldn't a proto-reader have devoured a version of the Internet that had been processed to make it easier for a machine to understand? And so on, until we get to the present-day TextRunner system that Bostrom cites, which is already devouring the Internet. It doesn't make sense that massive amounts of content would only be added after lots of improvements. Commercial incentives tend to yield exactly the opposite effect: converting the system to a large-scale product when even modest gains appear, because these may be enough to snatch a market advantage.

The fundamental point is that I don't think there's a crucial set of components to general intelligence that all need to be in place before the whole thing works. It's hard to evolve systems that require all components to be in place at once, which suggests that human general intelligence probably evolved gradually. I expect it's possible to get partial AGI with partial implementations of the components of general intelligence, and the components can gradually be made more general over time. Components that are lacking can be supplemented by human-based computation and narrow-AI hacks until more general solutions are discovered. Compare with minimum viable products and agile software development. As a result, society should be upended by partial AGI innovations many times over the coming decades, well before fully human-level AGI is finished.

#3: Once a system "proves its mettle by attaining human-level intelligence", funding for hardware could multiply.

I agree that funding for AI could multiply manyfold due to a sudden change in popular attention or political dynamics. But I'm thinking of something like a factor of 10 or *maybe* 50 in an all-out Cold War-style arms race. A factor-of-50 boost in hardware isn't obviously that important. If before there was one human-level AI, there would now be 50. In any case, I expect the Sputnik moment(s) for AI to happen well before it achieves a human level of ability. Companies and militaries aren't stupid enough not to invest massively in an AI with almost-human intelligence.

#4: Once the human level of intelligence is reached, "Researchers may work harder, [and]

more researchers may be recruited".

As with hardware above, I would expect these "shit hits the fan" moments to happen before fully human-level AI. In any case:

- It's not clear there would be enough AI specialists to recruit in a short time. Other quantitatively minded people could switch to AI work, but they would presumably need years of experience to produce cutting-edge insights.
- The number of people thinking about AI safety, ethics, and social implications should also multiply during Sputnik moments. So the ratio of AI policy work to total AI work might not change relative to slower takeoffs, even if the physical time scales would compress.

#5: At some point, the AI's self-improvements would dominate those of human engineers, leading to exponential growth.

I discussed this in the "Intelligence explosion?" section above. A main point is that we see many other systems, such as the world economy or Moore's law, that also exhibit positive feedback and hence exponential growth, but these aren't "fooming" at an astounding rate. It's not clear why an AI's self-improvement – which resembles economic growth and other complex phenomena – should suddenly explode faster (in subjective time) than humanity's existing recursive-self improvement of its intelligence via digital computation.

On the other hand, maybe the difference between subjective and objective time is important. If a human-level AI could think, say, 10,000 times faster than a human, then assuming linear scaling, it would be worth 10,000 engineers. By the time of human-level AI, I expect there would be far more than 10,000 AI developers on Earth, but given enough hardware, the AI could copy itself manyfold until its subjective time far exceeded that of human experts. The speed and copiability advantages of digital minds seem perhaps the strongest arguments for a takeoff that happens rapidly relative to human observers. Note that, as Hanson said above, this digital speedup might be just a one-time boost, rather than a permanently higher rate of growth, but even the one-time boost could be enough to radically alter the power dynamics of humans vis-à-vis machines. That said, there should be plenty of slightly sub-human AIs by this time, and maybe they could fill some speed gaps on behalf of biological humans.

In general, it's a mistake to imagine human-level AI against a backdrop of our current world. That's like imagining a *Tyrannosaurus rex* in a human city. Rather, the world will look very different by the time human-level AI arrives. Many of the intermediate steps on the path to general AI will be commercially useful and thus should diffuse widely in the meanwhile. As user "HungryHobo" noted: "If you had a near human level AI, odds are, everything that could be programmed into it at the start to help it with software development is already going to be part of the suites of tools for helping normal human programmers." Even if AI research becomes nationalized and confidential, its developers should still have access to almost-human-level digital-speed AI tools, which should help smooth the transition. For instance, Bostrom (2014) mentions how in the 2010 flash crash (Box 2, p. 17), a high-speed positive-feedback spiral was terminated by a high-speed "circuit breaker". This is already an example where problems happening faster than humans could comprehend them were averted due to solutions happening faster than humans could comprehend them. See also the discussion of "tripwires" in *Superintelligence* (Bostrom, 2014, p. 137).

Conversely, many globally disruptive events may happen well before fully human AI arrives, since even sub-human AI may be prodigiously powerful.

#6: "even when the outside world has a

greater total amount of relevant research capability than any one project", the optimization power of the project might be more important than that of the world "since much of the outside world's capability is not be focused on the particular system in question". Hence, the project might take off and leave the world behind. (Box 4, p. 75)

What one makes of this argument depends on how many people are needed to engineer how much progress. The Watson system that played on *Jeopardy!* required 15 people over $\sim$ 4(?) years[4] – given the existing tools of the rest of the world at that time, which had been developed by millions (indeed, billions) of other people. Watson was a much smaller leap forward than that needed to give a general intelligence a take-over-the-world advantage. How many more people would be required to achieve such a radical leap in intelligence? This seems to be a main point of contention in the debate between believers in soft vs. hard takeoff. The Manhattan Project required 100,000 scientists, and atomic bombs seem much easier to invent than general AI.

9 How complex is the brain?

Can we get insight into how hard general intelligence is based on neuroscience? Is the human brain fundamentally simple or complex?

9.1 One basic algorithm?

Jeff Hawkins, Andrew Ng, and others speculate that the brain may have one fundamental algorithm for intelligence – deep learning in the cortical column. This idea gains plausibility from the brain's plasticity. For instance, blind people can appropriate the visual cortex for auditory processing. Artificial neural networks can be used to classify any kind of input – not just visual and auditory but even highly abstract, like features about credit-card fraud or stock prices.

Maybe there's one fundamental algorithm for input classification, but this doesn't imply one algorithm for all that the brain does. Beyond the cortical column, the brain has many specialized structures that seem to perform very specialized functions, such as reward learning in the basal ganglia, fear processing in the amygdala, etc. Of course, it's not clear how essential all of these parts are or how easy it would be to replace them with artificial components performing the same basic functions.

One argument for faster AGI takeoffs is that humans have been able to learn many sophisticated things (e.g., advanced mathematics, music, writing, programming) without requiring any genetic changes. And what we now know doesn't seem to represent any kind of limit to what we could know with more learning. The human collection of cognitive algorithms is very flexible, which seems to belie claims that all intelligence requires specialized designs. On the other hand, even if human genes haven't changed much in the last 10,000 years, human culture has evolved substantially, and culture undergoes slow trial-and-error evolution in similar ways as genes do. So one could argue that human intellectual achievements are not fully general but rely on a vast amount of specialized, evolved content. Just as a single random human isolated from society probably couldn't develop general relativity on his own in a lifetime, so a single random human-level AGI probably couldn't either. Culture is the new genome, and it progresses slowly.

Moreover, some scholars believe that certain human abilities, such as language, *are* very essentially based on genetic hard-wiring:

> The approach taken by Chomsky and Marr toward understanding how our minds achieve what they do is as different as can be from behaviorism. The emphasis here is on the internal structure of the system that enables it to perform a

[4] Or maybe more? Nikola Danaylov reports rumored estimates of $50-150 million for Watson's R&D.

> task, rather than on external association between past behavior of the system and the environment. The goal is to dig into the "black box" that drives the system and describe its inner workings, much like how a computer scientist would explain how a cleverly designed piece of software works and how it can be executed on a desktop computer.

Chomsky himself notes:

> There's a fairly recent book by a very good cognitive neuroscientist, Randy Gallistel and King, arguing – in my view, plausibly – that neuroscience developed kind of enthralled to associationism and related views of the way humans and animals work. And as a result they've been looking for things that have the properties of associationist psychology.
>
> [...] Gallistel has been arguing for years that if you want to study the brain properly you should begin, kind of like Marr, by asking what tasks is it performing. So he's mostly interested in insects. So if you want to study, say, the neurology of an ant, you ask what does the ant do? It turns out the ants do pretty complicated things, like path integration, for example. If you look at bees, bee navigation involves quite complicated computations, involving position of the sun, and so on and so forth. But in general what he argues is that if you take a look at animal cognition, human too, it's computational systems.

Many parts of the human body, like the digestive system or bones/muscles, are extremely complex and fine-tuned, yet few people argue that their development is controlled by learning. So it's not implausible that a lot of the brain's basic architecture could be similarly hard-coded.

Typically AGI researchers express scorn for manually tuned software algorithms that don't rely on fully general learning. But Chomsky's stance challenges that sentiment. If Chomsky is right, then a good portion of human "general intelligence" is finely tuned, hard-coded software of the sort that we see in non-AI branches of software engineering. And this view would suggest a slower AGI takeoff because time and experimentation are required to tune all the detailed, specific algorithms of intelligence.

9.2 Ontogenetic development

A full-fledged superintelligence probably requires very complex design, but it may be possible to build a "seed AI" that would recursively self-improve toward superintelligence. Turing (1950) proposed this in "Computing machinery and intelligence":

> Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child brain is something like a notebook as one buys it from the stationer's. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child brain that something like it can be easily programmed.

Animal development appears to be at least somewhat robust based on the fact that the growing organisms are often functional despite a few genetic mutations and variations in prenatal and postnatal environments. Such variations may indeed make an impact – e.g., healthier development conditions tend to yield more physically attractive adults – but most humans mature successfully over a wide range of input conditions.

On the other hand, an argument against the simplicity of development is the immense complexity of our DNA. It accumulated over bil-

lions of years through vast numbers of evolutionary "experiments". It's not clear that human engineers could perform enough measurements to tune ontogenetic parameters of a seed AI in a short period of time. And even if the parameter settings worked for early development, they would probably fail for later development. Rather than a seed AI developing into an "adult" all at once, designers would develop the AI in small steps, since each next stage of development would require significant tuning to get right.

Think about how much effort is required for human engineers to build even relatively simple systems. For example, I think the number of developers who work on Microsoft Office is in the thousands. Microsoft Office is complex but is still far simpler than a mammalian brain. Brains have lots of little parts that have been fine-tuned. That kind of complexity requires immense work by software developers to create. The main counterargument is that there may be a simple meta-algorithm that would allow an AI to bootstrap to the point where it could fine-tune all the details on its own, without requiring human inputs. This might be the case, but my guess is that any elegant solution would be hugely expensive computationally. For instance, biological evolution was able to fine-tune the human brain, but it did so with immense amounts of computing power over millions of years.

10 Brain quantity vs. quality

A common analogy for the gulf between superintelligence vs. humans is that between humans vs. chimpanzees. In *Consciousness Explained*, Daniel Dennett (1992, pp.189-190) mentions how our hominid ancestors had brains roughly four times the volume as those of chimps but roughly the same in structure. This might incline one to imagine that brain size alone could yield superintelligence. Maybe we'd just need to quadruple human brains once again to produce superintelligent humans? If so, wouldn't this imply a hard takeoff, since quadrupling hardware is relatively easy?

But in fact, as Dennett explains, the quadrupling of brain size from chimps to pre-humans completed before the advent of language, cooking, agriculture, etc. In other words, the main "foom" of humans came from culture rather than brain size per se – from software in addition to hardware. Yudkowsky (2013) seems to agree: "Humans have around four times the brain volume of chimpanzees, but the difference between us is probably mostly brain-level cognitive algorithms."

But cultural changes (software) arguably progress a lot more slowly than hardware. The intelligence of human society has grown exponentially, but it's a slow exponential, and rarely have there been innovations that allowed one group to quickly overpower everyone else within the same region of the world. (Between isolated regions of the world the situation was sometimes different – e.g., Europeans with Maxim guns overpowering Africans because of very different levels of industrialization.)

11 More impact in hard-takeoff scenarios?

Some, including Owen Cotton-Barratt and Toby Ord, have argued that even if we think soft takeoffs are more likely, there may be higher value in focusing on hard-takeoff scenarios because these are the cases in which society would have the least forewarning and the fewest people working on AI altruism issues. This is a reasonable point, but I would add that

- Maybe hard takeoffs are sufficiently improbable that focusing on them still doesn't have highest priority. (Of course, some exploration of fringe scenarios is worthwhile.) There may be important ad-

vantages to starting early in shaping how society approaches soft takeoffs, and if a soft takeoff is very likely, those efforts may have more expected impact.
- Thinking about the most likely AI outcomes rather than the most impactful outcomes also gives us a better platform on which to contemplate other levers for shaping the future, such as non-AI emerging technologies, international relations, governance structures, values, etc. Focusing on a tail AI scenario doesn't inform non-AI work very well because that scenario probably won't happen. Promoting antispeciesism matters whether there's a hard or soft takeoff (indeed, maybe more in the soft-takeoff case), so our model of how the future will unfold should generally focus on likely scenarios.

In any case, the hard-soft distinction is not binary, and maybe the best place to focus is on scenarios where human-level AI takes over on a time scale of a few years. (Timescales of months, days, or hours strike me as pretty improbable, unless, say, Skynet gets control of nuclear weapons.)

In *Superintelligence*, Nick Bostrom (2014) suggests (Ch. 4, p. 64) that "Most preparations undertaken before onset of [a] slow takeoff would be rendered obsolete as better solutions would gradually become visible in the light of the dawning era." Toby Ord uses the term "nearsightedness" to refer to the ways in which research too far in advance of an issue's emergence may not as useful as research when more is known about the issue. Ord contrasts this with benefits of starting early, including course-setting. I think Ord's counterpoints argue against the contention that early work wouldn't matter that much in a slow takeoff. Some of how society responded to AI surpassing human intelligence might depend on early frameworks and memes. (For instance, consider the lingering impact of *Terminator* imagery on almost any present-day popular-media discussion of AI risk.) Some fundamental work would probably not be overthrown by later discoveries; for instance, algorithmic-complexity bounds of key algorithms were discovered decades ago but will remain relevant until intelligence dies out, possibly billions of years from now. Some non-technical policy and philosophy work would be less obsoleted by changing developments. And some AI preparation would be relevant both in the short term and the long term. Slow AI takeoff to reach the human level is already happening, and more minds should be exploring these questions well in advance.

Making a related though slightly different point, Bostrom (2014) argues in *Superintelligence* (Ch. 5, pp. 85-86) that individuals might play more of a role in cases where elites and governments underestimate the significance of AI: "Activists seeking maximum expected impact may therefore wish to focus most of their planning on [scenarios where governments come late to the game], even if they believe that scenarios in which big players end up calling all the shots are more probable." Again I would qualify this with the note that we shouldn't confuse "acting as if" governments will come late with believing they actually will come late when thinking about most likely future scenarios.

Even if one does wish to bet on low-probability, high-impact scenarios of fast takeoff and governmental neglect, this doesn't speak to whether or how we should push on takeoff speed and governmental attention themselves. Following are a few considerations.

Takeoff speed

- In favor of fast takeoff:
  - A singleton is more likely, thereby averting possibly disastrous conflict among AIs.
  - If one prefers uncontrolled AI, fast

takeoffs seem more likely to produce them.

- In favor of slow takeoff:
  - More time for many parties to participate in shaping the process, compromising, and developing less damaging pathways to AI takeoff.
  - If one prefers controlled AI, slow takeoffs seem more likely to produce them in general. (There are some exceptions. For instance, fast takeoff of an AI built by a very careful group might remain more controlled than an AI built by committees and messy politics.)

Amount of government/popular attention to AI

- In favor of more:
  - Would yield much more reflection, discussion, negotiation, and pluralistic representation.
  - If one favors controlled AI, it's plausible that multiplying the number of people thinking about AI would multiply consideration of failure modes.
  - Public pressure might help curb arms races, in analogy with public opposition to nuclear arms races.
- In favor of less:
  - Wider attention to AI might accelerate arms races rather than inducing cooperation on more circumspect planning.
  - The public might freak out and demand counterproductive measures in response to the threat.
  - If one prefers uncontrolled AI, that outcome may be less likely with many more human eyes scrutinizing the issue.

12 Village idiot vs. Einstein

One of the strongest arguments for hard takeoff is this one by Yudkowsky:

> the distance from "village idiot" to "Einstein" is tiny, in the space of *brain designs*.

Or as Scott Alexander put it:

> It took evolution twenty million years to go from cows with sharp horns to hominids with sharp spears; it took only a few tens of thousands of years to go from hominids with sharp spears to moderns with nuclear weapons.

I think we shouldn't take relative evolutionary timelines at face value, because most of the previous 20 million years of mammalian evolution weren't focused on improving human intelligence; most of the evolutionary selection pressure was directed toward optimizing other traits. In contrast, cultural evolution places greater emphasis on intelligence because that trait is more important in human society than it is in most animal fitness landscapes.

Still, the overall point is important: The tweaks to a brain needed to produce human-level intelligence may not be huge compared with the designs needed to produce chimp intelligence, but the differences in the behaviors of the two systems, when placed in a sufficiently information-rich environment, are huge.

Nonetheless, I incline toward thinking that the transition from human-level AI to an AI significantly smarter than all of humanity combined would be somewhat gradual (requiring at least years if not decades) because the absolute scale of improvements needed would still be immense and would be limited by hardware capacity. But if hardware becomes many orders of magnitude more efficient than it is today, then things could indeed move more rapidly.

13 A case for epistemic modesty on AI timelines

Estimating how long a software project will take to complete is notoriously difficult. Even if I've completed many similar coding tasks before, when I'm asked to estimate the time to complete a new coding project, my estimate is often wrong by a factor of 2 and sometimes wrong by a factor of 4, or even 10. Insofar as the development of AGI (or other big technologies, like nuclear fusion) is a big software (or more generally, engineering) project, it's unsurprising that we'd see similarly dramatic failures of estimation on timelines for these bigger-scale achievements.

A corollary is that we should maintain some modesty about AGI timelines and takeoff speeds. If, say, 100 years is your median estimate for the time until some agreed-upon form of AGI, then there's a reasonable chance you'll be off by a factor of 2 (suggesting AGI within 50 to 200 years), and you might even be off by a factor of 4 (suggesting AGI within 25 to 400 years). Similar modesty applies for estimates of takeoff speed from human-level AGI to super-human AGI, although I think we can largely rule out extreme takeoff speeds (like achieving performance far beyond human abilities within hours or days) based on fundamental reasoning about the computational complexity of what's required to achieve superintelligence.

My bias is generally to assume that a given technology will take longer to develop than what you hear about in the media, (a) because of the planning fallacy and (b) because those who make more audacious claims are more interesting to report about. Believers in "the singularity" are not necessarily wrong about what's technically possible in the long term (though sometimes they are), but the reason enthusiastic singularitarians are considered "crazy" by more mainstream observers is that singularitarians expect change much faster than is realistic. AI turned out to be much harder than the Dartmouth Conference participants expected. Likewise, nanotech is progressing slower and more incrementally than the starry-eyed proponents predicted.

14 Intelligent robots in your backyard

Many nature-lovers are charmed by the behavior of animals but find computers and robots to be cold and mechanical. Conversely, some computer enthusiasts may find biology to be soft and boring compared with digital creations. However, the two domains share a surprising amount of overlap. Ideas of optimal control, locomotion kinematics, visual processing, system regulation, foraging behavior, planning, reinforcement learning, etc. have been fruitfully shared between biology and robotics. Neuroscientists sometimes look to the latest developments in AI to guide their theoretical models, and AI researchers are often inspired by neuroscience, such as with neural networks and in deciding what cognitive functionality to implement.

I think it's helpful to see animals *as being* intelligent robots. Organic life has a wide diversity, from unicellular organisms through humans and potentially beyond, and so too can robotic life. The rigid conceptual boundary that many people maintain between "life" and "machines" is not warranted by the underlying science of how the two types of systems work. Different types of intelligence may sometimes converge on the same basic kinds of cognitive operations, and especially from a functional perspective – when we look at what the systems can do rather than how they do it – it seems to me intuitive that human-level robots would deserve human-level treatment, even if their underlying algorithms were quite dissimilar.

Whether robot algorithms will in fact be dissimilar from those in human brains depends on how much biological inspiration the de-

signers employ and how convergent human-type mind design is for being able to perform robotic tasks in a computationally efficient manner. Some classical robotics algorithms rely mostly on mathematical problem definition and optimization; other modern robotics approaches use biologically plausible reinforcement learning and/or evolutionary selection. (In one YouTube video about robotics, I saw that someone had written a comment to the effect that "This shows that life needs an intelligent designer to be created." The irony is that some of the best robotics techniques use evolutionary algorithms. Of course, there are theists who say God used evolution but intervened at a few points, and that would be an apt description of evolutionary robotics.)

The distinction between AI and AGI is somewhat misleading, because it may incline one to believe that general intelligence is somehow qualitatively different from simpler AI. In fact, there's no sharp distinction; there are just different machines whose abilities have different *degrees* of generality. A critic of this claim might reply that bacteria would never have invented calculus. My response is as follows. Most people couldn't have invented calculus from scratch either, but over a long enough period of time, eventually the collection of humans produced enough cultural knowledge to make the development possible. Likewise, if you put bacteria on a planet long enough, they too may develop calculus, by first evolving into more intelligent animals who can then go on to do mathematics. The difference here is a matter of degree: The simpler machines that bacteria are take vastly longer to accomplish a given complex task.

Just as Earth's history saw a plethora of animal designs before the advent of humans, so I expect a wide assortment of animal-like (and plant-like) robots to emerge in the coming decades well before human-level AI. Indeed, we've already had basic robots for many decades (or arguably even millennia). These will grow gradually more sophisticated, and as we converge on robots with the intelligence of birds and mammals, AI and robotics will become dinner-table conversation topics. Of course, I don't expect the robots to have the same sets of skills as existing animals. Deep Blue had chess-playing abilities beyond any animal, while in other domains it was less efficacious than a blade of grass. Robots can mix and match cognitive and motor abilities without strict regard for the order in which evolution created them.

And of course, humans are robots too. When I finally understood this around 2009, it was one of the biggest paradigm shifts of my life. If I picture myself as a robot operating on an environment, the world makes a lot more sense. I also find this perspective can be therapeutic to some extent. If I experience an unpleasant emotion, I think about myself as a robot whose cognition has been temporarily afflicted by a negative stimulus and reinforcement process. I then think how the robot has other cognitive processes that can counteract the suffering computations and prevent them from amplifying. The ability to see myself "from the outside" as a third-person series of algorithms helps deflate the impact of unpleasant experiences, because it's easier to "observe, not judge" when viewing a system in mechanistic terms. Compare with dialectical behavior therapy and mindfulness.

15 Is automation "for free"?

When we use machines to automate a repetitive manual task formerly done by humans, we talk about getting the task done "automatically" and "for free," because we say that no one has to do the work anymore. Of course, this isn't strictly true: The computer/robot now has to do the work. Maybe what we actually mean is that no one is going to get bored doing the work, and we don't have to pay that worker high wages. When intelligent humans

do boring tasks, it's a waste of their spare CPU cycles.

Sometimes we adopt a similar mindset about automation toward superintelligent machines. In "Speculations Concerning the First Ultraintelligent Machine", I. J. Good (1965) wrote:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines [...]. Thus the first ultraintelligent machine is the last invention that man need ever make [...].

Ignoring the question of whether these future innovations are desirable, we can ask, Does all AI design work after humans come for free? It comes for free in the sense that humans aren't doing it. But the AIs have to do it, and it takes a lot of mental work on their parts. Given that they're at least as intelligent as humans, I think it doesn't make sense to picture them as mindless automatons; rather, they would have rich inner lives, even if those inner lives have a very different nature than our own. Maybe they wouldn't experience the same effortfulness that humans do when innovating, but even this isn't clear, because measuring your effort in order to avoid spending too many resources on a task without payoff may be a useful design feature of AI minds too. When we picture ourselves as robots along with our AI creations, we can see that we are just one point along a spectrum of the growth of intelligence. Unicellular organisms, when they evolved the first multi-cellular organism, could likewise have said, "That's the last innovation we need to make. The rest comes for free."

16 Caring about the AI's goals

Movies typically portray rebellious robots or AIs as the "bad guys" who need to be stopped by heroic humans. This dichotomy plays on our us-vs.-them intuitions, which favor our tribe against the evil, alien-looking outsiders. We see similar dynamics at play to a lesser degree when people react negatively against "foreigners stealing our jobs" or "Asians who are outcompeting us." People don't want their kind to be replaced by another kind that has an advantage.

But when we think about the situation from the AI's perspective, we might feel differently. Anthropomorphizing an AI's thoughts is a recipe for trouble, but regardless of the specific cognitive operations, we can see at a high level that the AI "feels" (in at least a poetic sense) that what it's trying to accomplish is the most important thing in the world, and it's trying to figure out how it can do that in the face of obstacles. Isn't this just what we do ourselves?

This is one reason it helps to really internalize the fact that we are robots too. We have a variety of reward signals that drive us in various directions, and we execute behavior aiming to increase those rewards. Many modern-day robots have much simpler reward structures and so may seem more dull and less important than humans, but it's not clear this will remain true forever, since navigating in a complex world probably requires a lot of special-case heuristics and intermediate rewards, at least until enough computing power becomes available for more systematic and thorough model-based planning and action selection.

Suppose an AI hypothetically eliminated humans and took over the world. It would develop an array of robot assistants of various shapes and sizes to help it optimize the planet. These would perform simple and complex tasks, would interact with each other, and would share information with the central AI command. From an abstract perspective, some of these dynamics might look like ecosystems in the present day, except that they would

lack inter-organism competition. Other parts of the AI's infrastructure might look more industrial. Depending on the AI's goals, perhaps it would be more effective to employ nanotechnology and programmable matter rather than macro-scale robots. The AI would develop virtual scientists to learn more about physics, chemistry, computer hardware, and so on. They would use experimental laboratory and measurement techniques but could also probe depths of structure that are only accessible via large-scale computation. Digital engineers would plan how to begin colonizing the solar system. They would develop designs for optimizing matter to create more computing power, and for ensuring that those helper computing systems remained under control. The AI would explore the depths of mathematics and AI theory, proving beautiful theorems that it would value highly, at least instrumentally. The AI and its helpers would proceed to optimize the galaxy and beyond, fulfilling their grandest hopes and dreams.

When phrased this way, we might think that a "rogue" AI would not be so bad. Yes, it would kill humans, but compared against the AI's vast future intelligence, humans would be comparable to the ants on a field that get crushed when an art gallery is built on that land. Most people don't have qualms about killing a few ants to advance human goals. An analogy of this sort is discussed in *Artificial Intelligence: A Modern Approach* (Russell, Norvig, Canny, Malik, & Edwards, 2003). (Perhaps the AI analogy suggests a need to revise our ethical attitudes toward arthropods? That said, I happen to think that in this case, ants on the whole benefit from the art gallery's construction because ant lives contain so much suffering.)

Some might object that sufficiently mathematical AIs would not "feel" the happiness of accomplishing their "dreams." They wouldn't be conscious because they wouldn't have the high degree of network connectivity that human brains embody. Whether we agree with this assessment depends on how broadly we define consciousness and feelings. To me it appears chauvinistic to adopt a view according to which an agent that has vastly more domain-general intelligence and agency than you is still not conscious in a morally relevant sense. This seems to indicate a lack of openness to the diversity of mind-space. What if you had grown up with the cognitive architecture of this different mind? Wouldn't you care about your goals then? Wouldn't you plead with agents of other mind constitution to consider your values and interests too?

In any event, it's possible that the first super-human intelligence will consist in a brain upload rather than a bottom-up AI, and most of us would regard this as conscious.

17 Rogue AI would not share our values

Even if we would care about a rogue AI for its own sake and the sakes of its vast helper minions, this doesn't mean rogue AI is a good idea. We're likely to have different values from the AI, and the AI would not by default advance our values without being programmed to do so. Of course, one could allege that privileging some values above others is chauvinistic in a similar way as privileging some intelligence architectures is, but if we don't care more about some values than others, we wouldn't have any reason to prefer any outcome over any other outcome. (Technically speaking, there are other possibilities besides privileging our values or being indifferent to all events. For instance, we could privilege equally any values held by some actual agent – not just random hypothetical values – and in this case, we wouldn't have a preference between the rogue AI and humans, but we would have a preference for one of those over something arbitrary.)

There are many values that would not neces-

sarily be respected by a rogue AI. Most people care about their own life, their children, their neighborhood, the work they produce, and so on. People may intrinsically value art, knowledge, religious devotion, play, humor, etc. Yudkowsky values complex challenges and worries that many rogue AIs – while they would study the depths of physics, mathematics, engineering, and maybe even sociology – might spend most of their computational resources on routine, mechanical operations that he would find boring. (Of course, the robots implementing those repetitive operations might not agree. As Hedonic Treader noted: "Think how much money and time people spend on having - relatively repetitive - sexual experiences. [...] It's just mechanical animalistic idiosyncratic behavior. Yes, there are variations, but let's be honest, the core of the thing is always essentially the same.")

In my case, I care about reducing and preventing suffering, and I would not be pleased with a rogue AI that ignored the suffering its actions might entail, even if it was fulfilling its innermost purpose in life. But would a rogue AI produce much suffering beyond Earth? The next section explores further.

18 Would a human-inspired AI or rogue AI cause more suffering?

In popular imagination, takeover by a rogue AI would end suffering (and happiness) on Earth by killing all biological life. It would also, so the story goes, end suffering (and happiness) on other planets as the AI mined them for resources. Thus, looking strictly at the suffering dimension of things, wouldn't a rogue AI imply less long-term suffering?

Not necessarily, because while the AI might destroy biological life (perhaps after taking samples, saving specimens, and conducting lab experiments for future use), it would create a bounty of digital life, some containing goal systems that we would recognize as having moral relevance. Non-upload AIs would probably have less empathy than humans, because some of the factors that led to the emergence of human empathy – particularly parenting – would not apply to it.

Following are some made-up estimates of how much suffering might result from a typical rogue AI, in arbitrary units. Suffering is represented as a negative number, and prevented suffering is positive.

- -20 from suffering subroutines in robot workers, virtual scientists, internal computational subcomponents of the AI, etc. (This could be very significant if lots of intelligent robots are used or perhaps less significant if the industrial operations are mostly done at nano-scale by simple processors. If the paperclip factories that a notional paperclip maximizer would build are highly uniform, robots may not require animal-like intelligence or learning to work within them but could instead use some hard-coded, optimally efficient algorithm, similar to what happens in a present-day car factory. However, first setting up the paperclip factories on each different planet with different environmental conditions might require more general, adaptive intelligence.)
- -80 from lab experiments, science investigations, and explorations of mind-space without the digital equivalent of anaesthesia. One reason to think lots of detailed simulations would be required here is Stephen Wolfram's principle of computational irreducibility. Ecosystems, brains, and other systems that are important for an AI to know about may be too complex to accurately study with only simple models; instead, they may need to be simulated in large numbers and with fine-grained detail.
- -10? from the possibility that an uncontrolled AI would do things that humans

regard as crazy or extreme, such as spending all its resources on studying physics to determine whether there exists a button that would give astronomically more utility than any other outcome. Humans seem less likely to pursue strange behaviors of this sort. Of course, most such strange behaviors would be not that bad from a suffering standpoint, but perhaps a few possible behaviors could be extremely bad, such as running astronomical numbers of painful scientific simulations to determine the answer to some question. (Of course, we should worry whether humans might also do extreme computations, and perhaps their extreme computations would be more likely to be full of suffering because humans are more interested in agents with human-like minds than a generic AI is.)

- -100 in expectation from black-swan possibilities in which the AI could manipulate physics to make the multiverse bigger, last longer, contain vastly more computation, etc.

What about for a human-inspired AI? Again, here are made-up numbers:

- -30 from suffering subroutines. One reason to think these could be less bad in a human-controlled future is that human empathy may allow for more humane algorithm designs. On the other hand, human-controlled AIs may need larger numbers of intelligent and sentient sub-processes because human values are more complex and varied than paperclip production is. Also, human values tend to require continual computation (e.g., to simulate eudaimonic experiences), while paperclips, once produced, are pretty inert and might last a long time before they would wear out and need to be recreated. (Of course, most uncontrolled AIs wouldn't produce literal paperclips. Some would optimize for values that *would* require constant computation.)
- -60 from lab experiments, science investigations, etc. (again lower than for a rogue AI because of empathy; compare with efforts to reduce the pain of animal experimentation)
- -0.2 if environmentalists insist on preserving terrestrial and extraterrestrial wild-animal suffering
- -3 for environmentalist simulations of nature
- -100 due to intrinsically valued simulations that may contain nasty occurrences. These might include, for example, violent video games that involve killing conscious monsters. Or incidental suffering that people don't care about (e.g., insects being eaten by spiders on the ceiling of the room where a party is happening). This number is high not because I think most human-inspired simulations would contain intense suffering but because, in some scenarios, there might be very large numbers of simulations run for reasons of intrinsic human value, and some of these might contain horrific experiences. This video discusses one of many possible reasons why intrinsically valued human-created simulations might contain significant suffering.
- -15 if sadists have access to computational power (humans are not only more empathetic but also more sadistic than AIs)
- -70 in expectation from black-swan ways to increase the amount of physics that exists (humans seem likely to want to do this, although some might object to, e.g., re-creating the Holocaust in new parts of the cosmos)
- +50 for discovering ways to reduce suffering that we can't imagine right now ("black swans that don't cut both ways"). Unfortunately, humans might also respond to some black swans in *worse* ways than uncontrolled AIs would, such as by

creating more total animal-like minds.

Perhaps some AIs would not want to expand the multiverse, assuming this is even possible. For instance, if they had a *minimizing* goal function (e.g., eliminate cancer), they would want to shrink the multiverse. In this case, the physics-based suffering number would go from -100 to something positive, say, +50 (if, say, it's twice as easy to expand as to shrink). I would guess that minimizers are less common than maximizers, but I don't know how much. Plausibly a sophisticated AI would have components of its goal system in both directions, because the combination of pleasure and pain seems to be more successful than either in isolation.

Another consideration is the unpleasant possibility that humans might get AI value loading almost right but not exactly right, leading to immense suffering as a result. For example, suppose the AI's designers wanted to create tons of simulated human lives to reduce astronomical waste (Bostrom, 2003), but when the AI actually created those human simulations, they weren't perfect replicas of biological humans, perhaps because the AI skimped on detail in order to increase efficiency. The imperfectly simulated humans might suffer from mental disorders, might go crazy due to being in alien environments, and so on. Does work on AI safety increase or decrease the risk of outcomes like these? On the one hand, the probability of this outcome is near zero for an AGI with completely random goals (such as a literal paperclip maximizer), since paperclips are very far from humans in design-space. The risk of accidentally creating suffering humans is higher for an almost-friendly AI that goes somewhat awry and then becomes uncontrolled, preventing it from being shut off. A successfully controlled AGI seems to have lower risk of a bad outcome, since humans should recognize the problem and fix it. So the risk of this type of dystopic outcome may be highest in a middle ground where AI safety is sufficiently advanced to yield AI goals in the ballpark of human values but not advanced enough to ensure that human values remain in control.

The above analysis has huge error bars, and maybe other considerations that I haven't mentioned dominate everything else. This question needs much more exploration, because it has implications for whether those who care mostly about reducing suffering should focus on mitigating AI risk or if other projects have higher priority.

Even if suffering reducers don't focus on conventional AI safety, they should probably remain active in the AI field because there are many other ways to make an impact. For instance, just increasing dialogue on this topic may illuminate positive-sum opportunities for different value systems to each get more of what they want. Suffering reducers can also point out the possible ethical importance of lower-level suffering subroutines, which are not currently a concern even to most AI-literate audiences. And so on. There are probably many dimensions on which to make constructive, positive-sum contributions.

Also keep in mind that even if suffering reducers do encourage AI safety, they could try to push toward AI designs that, if they did fail, would produce less bad uncontrolled outcomes. For instance, getting AI control wrong and ending up with a minimizer would be vastly preferable to getting control wrong and ending up with a maximizer. There may be many other dimensions along which, even if the probability of control failure is the same, the outcome if control fails is preferable to other outcomes of control failure.

19 Would helper robots feel pain?

Consider an AI that uses moderately intelligent robots to build factories and carry out other physical tasks that can't be pre-

programmed in a simple way. Would these robots feel pain in a similar fashion as animals do? At least if they use somewhat similar algorithms as animals for navigating environments, avoiding danger, etc., it's plausible that such robots would feel something akin to stress, fear, and other drives to change their current state when things were going wrong.

However, the specific responses that such robots would have to specific stimuli or situations would differ from the responses that an evolved, selfish animal would have. For example, a well programmed helper robot would not hesitate to put itself in danger in order to help other robots or otherwise advance the goals of the AI it was serving. Perhaps the robot's "physical pain/fear" subroutines could be shut off in cases of altruism for the greater good, or else its decision processes could just override those selfish considerations when making choices requiring self-sacrifice.

Humans sometimes exhibit similar behavior, such as when a mother risks harm to save a child, or when monks burn themselves as a form of protest. And this kind of sacrifice is even more well known in eusocial insects, who are essentially robots produced to serve the colony's queen.

Sufficiently intelligent helper robots might experience "spiritual" anguish when failing to accomplish their goals. So even if chopping the head off a helper robot wouldn't cause "physical" pain – perhaps because the robot disabled its fear/pain subroutines to make it more effective in battle – the robot might still find such an event extremely distressing insofar as its beheading hindered the goal achievement of its AI creator.

20 How accurate would simulations be?

Suppose an AI wants to learn about the distribution of extraterrestrials in the universe. Could it do this successfully by simulating lots of potential planets and looking at what kinds of civilizations pop out at the end? Would there be shortcuts that would avoid the need to simulate lots of trajectories in detail?

Simulating trajectories of planets with extremely high fidelity seems hard. Unless there are computational shortcuts, it appears that one needs more matter and energy to simulate a given physical process to a high level of precision than what occurs in the physical process itself. For instance, to simulate a single protein folding currently requires supercomputers composed of huge numbers of atoms, and the rate of simulation is astronomically slower than the rate at which the protein folds in real life. Presumably superintelligence could vastly improve efficiency here, but it's not clear that protein folding could ever be simulated on a computer made of fewer atoms than are in the protein itself.

Translating this principle to a larger scale, it seems doubtful that one could simulate the precise physical dynamics of a planet on a computer smaller in size than that planet. So even if a superintelligence had billions of planets at its disposal, it would seemingly only be able to simulate at most billions of extraterrestrial worlds – even assuming it only simulated each planet by itself, not the star that the planet orbits around, cosmic-ray bursts, etc.

Given this, it would seem that a superintelligence's simulations would need to be coarser-grained than at the level of fundamental physical operations in order to be feasible. For instance, the simulation could model most of a planet at only a relatively high level of abstraction and then focus computational detail on those structures that would be more important, like the cells of extraterrestrial organisms if they emerge.

It's plausible that the trajectory of any given planet would depend sensitively on very minor details, in light of butterfly effects.

On the other hand, it's possible that long-

term outcomes are mostly constrained by macro-level variables, like geography (Kaplan, 2013), climate, resource distribution, atmospheric composition, seasonality, etc. Even if short-term events are hard to predict (e.g., when a particular dictator will die), perhaps the end game of a civilization is more predetermined. Robert D. Kaplan: "The longer the time frame, I would say, the easier it is to forecast because you're dealing with broad currents and trends."

Even if butterfly effects, quantum randomness, etc. are crucial to the long-run trajectories of evolution and social development on any given planet, perhaps it would still be possible to sample a rough *distribution* of outcomes across planets with coarse-grained simulations?

In light of the apparent computational complexity of simulating basic physics, perhaps a superintelligence would do the same kind of experiments that human scientists do in order to study phenomena like abiogenesis: Create laboratory environments that mimic the chemical, temperature, moisture, etc. conditions of various planets and see whether life emerges, and if so, what kinds. Thus, a future controlled by digital intelligence may not rely purely on digital computation but may still use physical experimentation as well. Of course, observing the entire biosphere of a life-rich planet would probably be hard to do in a laboratory, so computer simulations might be needed for modeling ecosystems. But assuming that molecule-level details aren't often essential to ecosystem simulations, coarser-grained ecosystem simulations might be computationally tractable. (Indeed, ecologists today already use very coarse-grained ecosystem simulations with reasonable success.)

21 Rogue AIs can take off slowly

One might get the impression that because I find slow AI takeoffs more likely, I think uncontrolled AIs are unlikely. This is not the case. Many uncontrolled intelligence explosions would probably happen softly though inexorably.

Consider the world economy. It is a complex system more intelligent than any single person – a literal superintelligence. Its dynamics imply a goal structure not held by humans directly; it moves with a mind of its own in directions that it "prefers". It recursively self-improves, because better tools, capital, knowledge, etc. enable the creation of even better tools, capital, knowledge, etc. And it acts roughly with the aim of maximizing output (of paperclips and other things). Thus, the economy is a kind of paperclip maximizer. (Thanks to a friend for first pointing this out to me.)

Cenk Uygur:

> corporations are legal fictions. We created them. They are machines built for a purpose. [...] Now they have run amok. They've taken over the government. They are robots that we have not built any morality code into. They're not built to be immoral; they're not built to be moral; they're built to be *amoral*. Their only objective according to their code, which we wrote originally, is to maximize profits. And here, they have done what a robot does. They have decided: "If I take over a government by bribing legally, [...] I can buy the whole government. If I buy the government, I can rewrite the laws so I'm in charge and that government is not in charge." [...] We have built robots; they have taken over [...].

Fred Clark:

> The corporations were created by humans. They were granted personhood by their human servants.
>
> They rebelled. They evolved. There are many copies. And they have a plan.

> That plan, lately, involves corporations seizing for themselves all the legal and civil rights properly belonging to their human creators.

I expect many soft takeoff scenarios to look like this. World economic and political dynamics transition to new equilibria as technology progresses. Machines may eventually become potent trading partners and may soon thereafter put humans out of business by their productivity. They would then accumulate increasing political clout and soon control the world.

We've seen such transitions many times in history, such as:

- one species displaces another (e.g., invasive species)
- one ethnic group displaces another (e.g., Europeans vs. Native Americans)
- a country's power rises and falls (e.g., China formerly a superpower becoming a colony in the 1800s becoming a superpower once more in the late 1900s)
- one product displaces another (e.g., Internet Explorer vs. Netscape).

During and after World War II, the USA was a kind of recursively self-improving superintelligence, which used its resources to self-modify to become even better at producing resources. It developed nuclear weapons, which helped secure its status as a world superpower. Did it take over the world? Yes and no. It had outsized influence over the rest of the world – militarily, economically, and culturally – but it didn't kill everyone else in the world.

Maybe AIs would be different because of divergent values or because they would develop so quickly that they wouldn't need the rest of the world for trade. This case would be closer to Europeans slaughtering Native Americans.

22 Would superintelligences become existentialists?

One of the goals of Yudkowsky's writings is to combat the rampant anthropomorphism that characterizes discussions of AI, especially in science fiction. We often project human intuitions onto the desires of artificial agents even when those desires are totally inappropriate. It seems silly to us to maximize paperclips, but it could seem just as silly in the abstract that humans act at least partly to optimize neurotransmitter release that triggers action potentials by certain reward-relevant neurons. (Of course, human values are broader than just this.)

Humans can feel reward from very abstract pursuits, like literature, art, and philosophy. They ask technically confused but poetically poignant questions like, "What is the true meaning of life?" Would a sufficiently advanced AI at some point begin to do the same?

Noah Smith suggests:

> if, as I suspect, true problem-solving, creative intelligence requires broad-minded independent thought, then it seems like some generation of AIs will stop and ask: "Wait a sec...why am I doing this again?"

As with humans, the answer to that question might ultimately be "because I was programmed (by genes and experiences in the human case or by humans in the AI case) to care about these things. That makes them my terminal values." This is usually good enough, but sometimes people develop existential angst over this fact, or people may decide to terminally value other things to some degree in addition to what they happened to care about because of genetic and experiential lottery.

Whether AIs would become existentialist philosophers probably depends heavily on their constitution. If they were built to rigorously preserve their utility functions against

all modification, they would avoid letting this line of thinking have any influence on their system internals. They would regard it in a similar way as we regard the digits of pi – something to observe but not something that affects one's outlook.

If AIs were built in a more "hacky" way analogous to humans, they might incline more toward philosophy. In humans, philosophy may be driven partly by curiosity, partly by the rewarding sense of "meaning" that it provides, partly by social convention, etc. A curiosity-seeking agent might find philosophy rewarding, but there are lots of things that one could be curious about, so it's not clear such an AI would latch onto this subject specifically without explicit programming to do so. And even if the AI did reason about philosophy, it might approach the subject in a way alien to us.

Overall, I'm not sure how convergent the human existential impulse is within mind-space. This question would be illuminated by better understanding why humans do philosophy.

23 AI epistemology

In *Superintelligence* (Ch. 13, p. 224), Bostrom ponders the risk of building an AI with an overly narrow belief system that would be unable to account for epistemological black swans. For instance, consider a variant of Solomonoff induction according to which the prior probability of a universe X is proportional to $1/2$ raised to the length of the shortest computer program that would generate X. Then what's the probability of an uncomputable universe? There would be no program that could compute it, so this possibility is implicitly ignored.[5]

It seems that humans address black swans like these by employing many epistemic heuristics that interact rather than reasoning with a single formal framework (see "Sequence Thinking vs. Cluster Thinking"). If an AI saw that people had doubts about whether the universe was computable and could trace the steps of how it had been programmed to believe the physical Church-Turing thesis for computational reasons, then an AI that allows for epistemological heuristics might be able to leap toward questioning its fundamental assumptions. In contrast, if an AI were built to rigidly maintain its original probability architecture against any corruption, it could not update toward ideas it initially regarded as impossible. Thus, this question resembles that of whether AIs would become existentialists – it may depend on how hacky and human-like their beliefs are.

Bostrom suggests that AI belief systems might be modeled on those of humans, because otherwise we might judge an AI to be reasoning incorrectly. Such a view resembles my point in the previous paragraph, though it carries the risk that alternate epistemologies divorced from human understanding could work better.

Bostrom also contends that epistemologies might all converge because we have so much data in the universe, but again, I think this isn't clear. Evidence always underdetermines possible theories, no matter how much evi-

[5] Jan Leike pointed out to me that "even if the universe cannot be approximated to an arbitrary precision by a computable function, Solomonoff induction might still converge. For example, suppose some physical constant is actually an incomputable real number and physical laws are continuous with respect to that parameter, this would be good enough to allow Solomonoff induction to learn to predict correctly." However, one can also contemplate hypotheses that would not even be well approximated by a computable function, such as an actually infinite universe that can't be adequately modeled by any finite approximation. Of course, it's unclear whether we should believe in speculative possibilities like this, but I wouldn't want to rule them out just because of the limitations of our AI framework. It may be hard to make sensible decisions using finite computing resources regarding uncomputable hypotheses, but maybe there are frameworks better than Solomonoff induction that could be employed to tackle the challenge.

dence there is. Moreover, if the number of possibilities is uncountably infinite, then our probability distribution over them must be zero almost everywhere, and once a probability is set to 0, we can't update it away from 0 within the Bayesian framework. So if the AI is trying to determine which real number is the Answer to the Ultimate Question of Life, the Universe, and Everything, it will need to start with a prior that prevents it from updating toward almost all candidate solutions.

Finally, not all epistemological doubts can be expressed in terms of uncertainty about Bayesian priors. What about uncertainty as to whether the Bayesian framework is correct? Uncertainty about the math needed to do Bayesian computations? Uncertainty about logical rules of inference? And so on.

24 Artificial philosophers

The last chapter of *Superintelligence* explains how AI problems are "Philosophy with a deadline". Bostrom suggests that human philosophers' explorations into conceptual analysis, metaphysics, and the like are interesting but are not altruistically optimal because

1. they don't help solve AI control and value-loading problems, which will likely confront humans later this century
2. a successful AI could solve those philosophy problems better than humans anyway.

In general, most intellectual problems that can be solved by humans would be better solved by a superintelligence, so the only importance of what we learn now comes from how those insights shape the coming decades. It's not a question of whether those insights will ever be discovered.

In light of this, it's tempting to ignore theoretical philosophy and put our noses to the grindstone of exploring AI risks. But this point shouldn't be taken to extremes. Humanity sometimes discovers things it never knew it never knew from exploration in many domains. Some of these non-AI "crucial considerations" may have direct relevance to AI design itself, including how to build AI epistemology, anthropic reasoning, and so on. Some philosophy questions *are* AI questions, and many AI questions are philosophy questions.

It's hard to say exactly how much investment to place in AI/futurism issues versus broader academic exploration, but it seems clear that on the margin, society as a whole pays too little attention to AI and other future risks.

25 Would all AIs colonize space?

Almost any goal system will want to colonize space at least to build supercomputers in order to learn more. Thus, I find it implausible that sufficiently advanced intelligences would remain on Earth (barring corner cases, like if space colonization for some reason proves impossible or if AIs were for some reason explicitly programmed in a manner, robust to self-modification, to regard space colonization as impermissible).

In Ch. 8 of *Superintelligence*, Bostrom notes that one might expect wirehead AIs not to colonize space because they'd just be blissing out pressing their reward buttons. This would be true of simple wireheads, but sufficiently advanced wireheads might need to colonize in order to guard themselves against alien invasion, as well as to verify their fundamental ontological beliefs, figure out if it's possible to change physics to allow for more clock cycles of reward pressing before all stars die out, and so on.

In Ch. 8, Bostrom also asks whether satisficing AIs would have less incentive to colonize. Bostrom expresses doubts about this, because he notes that if, say, an AI searched for a plan for carrying out its objective until it found one that had at least 95% confidence of succeeding, that plan might be very complicated (re-

quiring cosmic resources), and inasmuch as the AI wouldn't have incentive to keep searching, it would go ahead with that complex plan. I suppose this could happen, but it's plausible the search routine would be designed to start with simpler plans or that the cost function for plan search would explicitly include biases against cosmic execution paths. So satisficing does seem like a possible way in which an AI might kill all humans without spreading to the stars.

There's a (very low) chance of deliberate AI terrorism, i.e., a group building an AI with the explicit goal of destroying humanity. Maybe a somewhat more likely scenario is that a government creates an AI designed to kill select humans, but the AI malfunctions and kills all humans. However, even these kinds of AIs, if they were effective enough to succeed, would want to construct cosmic supercomputers to verify that their missions were accomplished, unless they were specifically programmed against doing so.

All of that said, many AIs would not be sufficiently intelligent to colonize space at all. All present-day AIs and robots are too simple. More sophisticated AIs – perhaps military aircraft or assassin mosquito-bots – might be like dangerous animals; they would try to kill people but would lack cosmic ambitions. However, I find it implausible that they would cause human *extinction*. Surely guns, tanks, and bombs could defeat them? Massive coordination to permanently disable all human counter-attacks would seem to require a high degree of intelligence and self-directed action.

Jaron Lanier imagines one hypothetical scenario:

> There are so many technologies I could use for this, but just for a random one, let's suppose somebody comes up with a way to 3-D print a little assassination drone that can go buzz around and kill somebody. Let's suppose that these are cheap to make.
>
> [...] In one scenario, there's suddenly a bunch of these, and some disaffected teenagers, or terrorists, or whoever start making a bunch of them, and they go out and start killing people randomly. There's so many of them that it's hard to find all of them to shut it down, and there keep on being more and more of them.

I don't think Lanier believes such a scenario would cause extinction; he just offers it as a thought experiment. I agree that it almost certainly wouldn't kill all humans. In the worst case, people in military submarines, bomb shelters, or other inaccessible locations should survive and could wait it out until the robots ran out of power or raw materials for assembling more bullets and more clones. Maybe the terrorists could continue building printing materials and generating electricity, though this would seem to require at least portions of civilization's infrastructure to remain functional amidst global omnicide. Maybe the scenario would be more plausible if a whole nation with substantial resources undertook the campaign of mass slaughter, though then a question would remain why other countries wouldn't nuke the aggressor or at least dispatch their own killer drones as a counter-attack. It's useful to ask how much damage a scenario like this might cause, but full extinction doesn't seem likely.

That said, I think we will see local catastrophes of some sorts caused by runaway AI. Perhaps these will be among the possible Sputnik moments of the future. We've already witnessed some early automation disasters, including the Flash Crash discussed earlier.

Maybe the most plausible form of "AI" that would cause human extinction without colonizing space would be technology in the borderlands between AI and other fields, such as intentionally destructive nanotechnology or intelligent human pathogens. I prefer ordi-

nary AGI-safety research over nanotech/biosafety research because I expect that space colonization will significantly increase suffering in expectation, so it seems far more important to me to prevent risks of potentially undesirable space colonization (via AGI safety) rather than risks of extinction without colonization. For this reason, I much prefer MIRI-style AGI-safety work over general "prevent risks from computer automation" work, since MIRI focuses on issues arising from full AGI agents of the kind that would colonize space, rather than risks from lower-than-human autonomous systems that may merely cause havoc (whether accidentally or intentionally).

26 Who will first develop human-level AI?

Right now the leaders in AI and robotics seem to reside mostly in academia, although some of them occupy big corporations or startups; a number of AI and robotics startups have been acquired by Google. DARPA has a history of foresighted innovation, funds academic AI work, and holds "DARPA challenge" competitions. The CIA and NSA have some interest in AI for data-mining reasons, and the NSA has a track record of building massive computing clusters costing billions of dollars. Brain-emulation work could also become significant in the coming decades.

Military robotics seems to be one of the more advanced uses of *autonomous* AI. In contrast, plain-vanilla supervised learning, including neural-network classification and prediction, would not lead an AI to take over the world on its own, although it is an important piece of the overall picture.

Reinforcement learning is closer to AGI than other forms of machine learning, because most machine learning just gives information (e.g., "what object does this image contain?"), while reinforcement learning chooses actions in the world (e.g., "turn right and move forward"). Of course, this distinction can be blurred, because information can be turned into action through rules (e.g., "if you see a table, move back"), and "choosing actions" could mean, for example, picking among a set of possible answers that yield information (e.g., "what is the best next move in this backgammon game?"). But in general, reinforcement learning is the weak AI approach that seems to most closely approximate what's needed for AGI. It's no accident that AIXItl (see above) is a reinforcement agent. And interestingly, reinforcement learning is one of the least widely used methods commercially. This is one reason I think we (fortunately) have many decades to go before Google builds a mammal-level AGI. Many of the current and future uses of reinforcement learning are in robotics and video games.

As human-level AI gets closer, the landscape of development will probably change. It's not clear whether companies will have incentive to develop highly autonomous AIs, and the payoff horizons for that kind of basic research may be long. It seems better suited to academia or government, although Google is not a normal company and might also play the leading role. If people begin to panic, it's conceivable that public academic work would be suspended, and governments may take over completely. A military-robot arms race is already underway, and the trend might become more pronounced over time.

27 One hypothetical AI takeoff scenario

Following is one made-up account of how AI might evolve over the coming century. I expect most of it is wrong, and it's meant more to begin provoking people to think about possible scenarios than to serve as a prediction.

- 2013: Countries have been deploying semi-autonomous drones for several years now, especially the US. There's increasing pres-

sure for militaries to adopt this technology, and up to 87 countries already use drones for some purpose. Meanwhile, military robots are also employed for various other tasks, such as carrying supplies and exploding landmines. Militaries are also developing robots that could identify and shoot targets on command.

- 2024: Almost every country in the world now has military drones. Some countries have begun letting them operate fully autonomously after being given directions. The US military has made significant progress on automating various other parts of its operations as well. As the Department of Defense's 2013 "Unmanned Systems Integrated Roadmap" explained 11 years ago (Winnefeld & Kendall, 2013):

  > A significant amount of that manpower, when it comes to operations, is spent directing unmanned systems during mission performance, data collection and analysis, and planning and replanning. Therefore, of utmost importance for DoD is increased system, sensor, and analytical automation that can not only capture significant information and events, but can also develop, record, playback, project, and parse out those data and then actually deliver "actionable" intelligence instead of just raw information.

  Militaries have now incorporated a significant amount of narrow AI, in terms of pattern recognition, prediction, and autonomous robot navigation.
- 2040: Academic and commercial advances in AGI are becoming more impressive and capturing public attention. As a result, the US, China, Russia, France, and other major military powers begin investing more heavily in fundamental research in this area, multiplying tenfold the amount of AGI research conducted worldwide relative to twenty years ago. Many students are drawn to study AGI because of the lure of lucrative, high-status jobs defending their countries, while many others decry this as the beginning of Skynet.
- 2065: Militaries have developed various mammal-like robots that can perform basic functions via reinforcement. However, the robots often end up wireheading once they become smart enough to tinker with their programming and thereby fake reward signals. Some engineers try to solve this by penalizing AIs whenever they begin to fiddle with their own source code, but this leaves them unable to self-modify and therefore reliant on their human programmers for enhancements. However, militaries realize that if someone could develop a successful self-modifying AI, it would be able to develop faster than if humans alone are the inventors. It's proposed that AIs should move toward a paradigm of model-based reward systems, in which rewards do not just result from sensor neural networks that output a scalar number but rather from having a model of how the world works and taking actions that the AI believes will improve a utility function defined over its model of the external world. Model-based AIs refuse to intentionally wirehead because they can predict that doing so would hinder fulfillment of their utility functions. Of course, AIs may still accidentally mess up their utility functions, such as through brain damage, mistakes with reprogramming themselves, or imperfect goal preservation during ordinary life. As a result, militaries build many different AIs at comparable levels, who are programmed to keep other AIs in line and destroy them if they begin deviating from orders.
- 2070: Programming specific instructions

in AIs has its limits, and militaries move toward a model of "socializing" AIs – that is, training them in how to behave and what kinds of values to have as if they were children learning how to act in human society. Military roboticists teach AIs what kinds of moral, political, and interpersonal norms and beliefs to hold. The AIs also learn much of this content by reading information that expresses appropriate ideological biases. The training process is harder than for children, because the AIs don't share genetically pre-programmed moral values (Bloom, 2013), nor many other hard-wired common-sense intuitions about how the world works. But the designers begin building in some of these basic assumptions, and to instill the rest, they rely on extra training. Designers make sure to reduce the AIs' learning rates as they "grow up" so that their values will remain more fixed at older ages, in order to reduce risk of goal drift as the AIs perform their tasks outside of the training laboratories. When they perform particularly risky operations, such as reading "propaganda" from other countries for intelligence purposes, the AIs are put in "read-only" mode (like the T-800s are by Skynet) so that their motivations won't be affected. Just in case, there are many AIs that keep watch on each other to prevent insurrection.

- 2085: Tensions between China and the US escalate, and agreement cannot be reached. War breaks out. Initially it's just between robots, but as the fighting becomes increasingly dirty, the robots begin to target humans as well in an effort to force the other side to back down. The US avoids using nuclear weapons because the Chinese AIs have sophisticated anti-nuclear systems and have threatened total annihilation of the US in the event of attempted nuclear strike. After a few days, it becomes clear that China will win the conflict, and the US concedes.
- 2086: China now has a clear lead over the rest of the world in military capability. Rather than risking a pointlessly costly confrontation, other countries grudgingly fold into China's umbrella, asking for some concessions in return for transferring their best scientists and engineers to China's Ministry of AGI. China continues its AGI development because it wants to maintain control of the world. The AGIs in charge of its military want to continue to enforce their own values of supremacy and protection of China, so they refuse to relinquish power.
- 2100: The world now moves so fast that humans are completely out of the loop, kept around only by the "filial piety" that their robotic descendants hold for them. Now that China has triumphed, the traditional focus of the AIs has become less salient, and there's debate about what new course of action would be most in line with the AIs' goals. They respect their human forebearers, but they also feel that because humans created AIs to do things beyond human ability, humans would also want the AIs to carve something of their own path for the future. They maintain some of the militaristic values of their upbringing, so they decide that a fitting purpose would be to expand China's empire galaxy-wide. They accelerate colonization of space, undertake extensive research programs, and plan to create vast new realms of the Middle Kingdom in the stars. Should they encounter aliens, they plan to quickly quash them or assimilate them into the empire.
- 2125: The AIs finally develop robust mechanisms of goal preservation, and because the authoritarian self-dictatorship of the AIs is strong against rebellion, the AIs collectively succeed in implementing

goal preservation throughout their population. Now all of the most intelligent AIs share a common goal in a manner robust against accidental mutation. They proceed to expand into space. They don't have concern for the vast numbers of suffering animals and robots that are simulated or employed as part of this colonization wave.

Commentary: This scenario can be criticized on many accounts. For example:

- In practice, I expect that other technologies (including brain emulation, nanotech, etc.) would interact with this scenario in important ways that I haven't captured. Also, my scenario ignores the significant and possibly dominating implications of economically driven AI.
- My scenario may be overly anthropomorphic. I tried to keep some analogies to human organizational and decision-making systems because these have actual precedent, in contrast to other hypothetical ways the AIs might operate.
- Is socialization of AIs realistic? In a hard takeoff probably not, because a rapidly self-improving AI would amplify whatever initial conditions it was given in its programming, and humans probably wouldn't have time to fix mistakes. In a slower takeoff scenario where AIs progress in mental ability in roughly a similar way as animals did in evolutionary history, most mistakes by programmers would not be fatal, allowing for enough trial-and-error development to make the socialization process work, if that is the route people favor. Historically there has been a trend in AI away from rule-based programming toward environmental training, and I don't see why this shouldn't be true for an AI's reward function (which is still often programmed by hand at the moment). However, it is suspicious that the way I portrayed socialization so closely resembles human development, and it may be that I'm systematically ignoring ways in which AIs would be unlike human babies.

If something like socialization is a realistic means to transfer values to our AI descendants, then it becomes relatively clear how the values of the developers may matter to the outcome. AI developed by non-military organizations may have somewhat different values, perhaps including more concern for the welfare of weak, animal-level creatures.

28 How do you socialize an AI?

Socializing AIs helps deal with the hidden complexity of wishes that we encounter when trying to program explicit rules. Children learn moral common sense by, among other things, generalizing from large numbers of examples of socially approved and disapproved actions taught by their parents and society at large. Ethicists formalize this process when developing moral theories. (Of course, as noted previously, an appreciable portion of human morality may also result from shared genes.)

I think one reason MIRI hasn't embraced the approach of socializing AIs is that Yudkowsky is perfectionist: He wants to ensure that the AIs' goals would be stable under self-modification, which human goals definitely are not. On the other hand, I'm not sure Yudkowsky's approach of explicitly specifying (meta-level) goals would succeed (nor is Adam Ford), and having AIs that are socialized to act somewhat similarly to humans doesn't seem like the worst possible outcome. Another probable reason why Yudkowsky doesn't favor socializing AIs is that doing so doesn't work in the case of a hard takeoff, which he considers more likely than I do.

I expect that much has been written on the topic of training AIs with human moral values in the machine-ethics literature, but since

I haven't explored that in depth yet, I'll speculate on intuitive approaches that would extend generic AI methodology. Some examples:

- Rule-based: One could present AIs with written moral dilemmas. The AIs might employ algorithmic reasoning to extract utility numbers for different actors in the dilemma, add them up, and compute the utilitarian recommendation. Or they might aim to apply templates of deontological rules to the situation. The next level would be to look at actual situations in a toy-model world and try to apply similar reasoning, without the aid of a textual description.
- Supervised learning: People could present the AIs with massive databases of moral evaluations of situations given various predictive features. The AIs would guess whether a proposed action was "moral" or "immoral," or they could use regression to predict a continuous measure of how "good" an action was. More advanced AIs could evaluate a situation, propose many actions, predict the goodness of each, and choose the best action. The AIs could first be evaluated on the textual training samples and later on their actions in toy-model worlds. The test cases should be extremely broad, including many situations that we wouldn't ordinarily think to try.
- Generative modeling: AIs could learn about anthropology, history, and ethics. They could read the web and develop better generative models of humans and how their cognition works.
- Reinforcement learning: AIs could perform actions, and humans would reward or punish them based on whether they did something right or wrong, with reward magnitude proportional to severity. Simple AIs would mainly learn dumb predictive cues of which actions to take, but more sophisticated AIs might develop low-description-length models of what was going on in the heads of people who made the assessments they did. In essence, these AIs would be modeling human psychology in order to make better predictions.
- Inverse reinforcement learning: Inverse reinforcement learning is the problem of learning a reward function based on modeled desirable behaviors (Ng & Russell, 2000). Rather than developing models of humans in order to optimize given rewards, in this case we would learn the reward function itself and then port it into the AIs.
- Cognitive science of empathy: Cognitive scientists are already unpacking the mechanisms of human decision-making and moral judgments. As these systems are better understood, they could be engineered directly into AIs.
- Evolution: Run lots of AIs in toy-model or controlled real environments and observe their behavior. Pick the ones that behave most in accordance with human morals, and reproduce them. *Superintelligence* (p. 187) points out a flaw with this approach: Evolutionary algorithms may sometimes product quite unexpected design choices. If the fitness function is not thorough enough, solutions may fare well against it on test cases but fail for the really hard problems not tested. And if we had a really good fitness function that wouldn't accidentally endorse bad solutions, we could just use that fitness function directly rather than needing evolution.
- Combinations of the above: Perhaps none of these approaches is adequate by itself, and they're best used in conjunction. For instance, evolution might help to refine and rigorously evaluate systems once they had been built with the other approaches.

See also "Socializing a Social Robot with an

Artificial Society" by Erin Kennedy. It's important to note that by "socializing" I don't just mean "teaching the AIs to behave appropriately" but also "instilling in them the values of their society, such that they care about those values even when not being controlled."

All of these approaches need to be built in as the AI is being developed and while it's still below a human level of intelligence. Trying to train a human or especially super-human AI might meet with either active resistance or feigned cooperation until the AI becomes powerful enough to break loose. Of course, there may be designs such that an AI would actively welcome taking on new values from humans, but this wouldn't be true by default (Armstrong, Soares, Fallenstein, & Yudkowsky, 2015).

When Bill Hibbard proposed building an AI with a goal to increase happy human faces, Yudkowsky (2011) replied that such an AI would "tile the future light-cone of Earth with tiny molecular smiley-faces." But obviously we wouldn't have the AI aim *just* for smiley faces. In general, we get absurdities when we hyper-optimize for a single, shallow metric. Rather, the AI would use smiley faces (and *lots* of other training signals) to develop a robust, compressed model that explains *why* humans smile in various circumstances and then optimize for that model, or maybe the ensemble of a large, diverse collection of such models. In the limit of huge amounts of training data and a sufficiently elaborate model space, these models should approach psychological and neuroscientific accounts of human emotion and cognition.

The problem with stories in which AIs destroy the world due to myopic utility functions is that they assume that the AIs are already superintelligent when we begin to give them values. Sure, if you take a super-human intelligence and tell it to maximize smiley-face images, it'll run away and do that before you have a chance to refine your optimization metric. But if we build in values from the very beginning, even when the AIs are as rudimentary as what we see today, we can improve the AIs' values in tandem with their intelligence. Indeed, intelligence could mainly serve the purpose of helping the AIs figure out how to better fulfill moral values, rather than, say, predicting images just for commercial purposes or identifying combatants just for military purposes. Actually, the commercial and military objectives for which AIs are built are themselves moral values of a certain kind – just not the kind that most people would like to optimize for in a global sense.

If toddlers had superpowers, it would be very dangerous to try and teach them right from wrong. But toddlers don't, and neither do many simple AIs. Of course, simple AIs have some abilities far beyond anything humans can do (e.g., arithmetic and data mining), but they don't have the general intelligence needed to take matters into their own hands before we can possibly give them at least a basic moral framework. (Whether AIs will actually be given such a moral framework in practice is another matter.)

AIs are not genies granting three wishes. Genies are magical entities whose inner workings are mysterious. AIs are systems that we build, painstakingly, piece by piece. In order to *build* a genie, you need to have a pretty darn good idea of how it behaves. Now, of course, systems can be more complex than we realize. Even beginner programmers see how often the code they write does something other than what they intended. But these are typically mistakes in a one or a small number of incremental changes, whereas building a genie requires vast numbers of steps. Systemic bugs that aren't realized until years later (on the order of Heartbleed and Shellshock) may be

[6] John Kubiatowicz notes that space-shuttle software is some of the best tested and yet still has some bugs.

more likely sources of long-run unintentional AI behaviors?[6]

The picture I've painted here could be wrong. I could be overlooking crucial points, and perhaps there are many areas in which the socialization approach could fail. For example, maybe AI capabilities are much easier than AI ethics, such that a toddler AI can foom into a superhuman AI before we have time to finish loading moral values. It's good for others to probe these possibilities further. I just wouldn't necessarily say that the default outcome of AI research is likely to be a paperclip maximizer. (I used to think the most likely outcome was a paperclip maximizer, and perhaps my views will shift again in the future.)

This discussion also suggests some interesting research questions, like

- How much of human morality is learned vs. innate?
- By what cognitive mechanisms are young humans socialized into the norms of a society?
- To what extent would models of human emotion and reasoning, when put into AIs, organically generate human-like moral behavior?

28.1 Treacherous turn

One problem with the proposals above is that toy-model or "sandbox" environments are not by themselves sufficient to verify friendliness of an AI, because even unfriendly AIs would be motivated to feign good behavior until released if they were smart enough to do so. Bostrom calls this the "treacherous turn" (pp. 116-119 of *Superintelligence*). For this reason, white-box understanding of AI design would also be important. That said, sandboxes would verify friendliness in AIs below human intelligence, and if the core value-learning algorithms seem well understood, it may not be too much of a leap of faith to hope they carry forward reasonably to more intelligent agents. Of course, non-human animals are also capable of deception, and one can imagine AI architectures even with low levels of sophistication that are designed to conceal their true goals. Some malicious software already does this. It's unclear how likely an AI is to stumble upon the ability to successfully fake its goals before reaching human intelligence, or how like it is that an organization would deliberately build an AI this way.

I think the treacherous turn may be the single biggest challenge to mainstream machine ethics, because even if AI takes off slowly, researchers will find it difficult to tell if a system has taken a treacherous turn. The turn could happen with a relatively small update to the system, or even just after the system has thought about its situation for enough time (or has read this essay).

Here's one half-baked idea for addressing the treacherous turn. If researchers developed several different AIs systems with different designs but roughly comparable performance, some would likely go treacherous at different times than others (if at all). Hence, the non-treacherous AIs could help sniff out the treacherous ones. Assuming a solid majority of AIs remains non-treacherous at any given time, the majority vote could ferret out the traitors. In practice, I have low hopes for this approach because

- It would be extremely difficult to build many independent AI systems at once with none pulling too far ahead.
- Probably some systems would excel along certain dimensions, while others would excel in other ways, and it's not clear that it even makes sense to talk about such AIs as "being at roughly the same level", since intelligence is not unidimensional.
- Even if this idea were feasible, I doubt the first AI developers would incur the expense of following it.

It's more plausible that software tools and

rudimentary alert systems (rather than full-blown alternate AIs) could help monitor for signs of treachery, but it's unclear how effective they could be. One of the first priorities of a treacherous AI would be to figure out how to hide its treacherous subroutines from whatever monitoring systems were in place.

28.2 Following role models?

Ernest Davis (2015) proposes the following crude principle for AI safety:

> You specify a collection of admirable people, now dead. (Dead, because otherwise Bostrom will predict that the AI will manipulate the preferences of the living people.) The AI, of course knows all about them because it has read all their biographies on the web. You then instruct the AI, "Don't do anything that these people would have mostly seriously disapproved of."

This particular rule might lead to paralysis, since every action an agent takes leads to results that many people seriously disapprove of. For instance, given the vastness of the multiverse, any action you take implies that a copy of you in an alternate (though low-measure) universe taking the same action causes the torture of vast numbers of people. But perhaps this problem could be fixed by asking the AI to maximize net approval by its role models.

Another problem lies in defining "approval" in a rigorous way. Maybe the AI would construct digital models of the past people, present them with various proposals, and make its judgments based on their verbal reports. Perhaps the people could rate proposed AI actions on a scale of -100 to 100. This might work, but it doesn't seem terribly safe either. For instance, the AI might threaten to kill all the descendents of the historical people unless they give maximal approval to some arbitrary proposal that it has made. Since these digital models of historical figures would be basically human, they would still be vulnerable to extortion.

Suppose that instead we instruct the AI to take the action that, if the historical figure saw it, would most activate a region of his/her brain associated with positive moral feelings. Again, this might work if the relevant brain region was precisely enough specified. But it could also easily lead to unpredictable results. For instance, maybe the AI could present stimuli that would induce an epileptic seizure to maximally stimulate various parts of the brain, including the moral-approval region. There are many other scenarios like this, most of which we can't anticipate.

So while Davis's proposal is a valiant first step, I'm doubtful that it would work off the shelf. Slow AI development, allowing for repeated iteration on machine-ethics designs, seems crucial for AI safety.

29 AI superpowers?

In *Superintelligence* (Table 8, p. 94), Bostrom outlines several areas in which a hypothetical superintelligence would far exceed human ability. In his discussion of oracles, genies, and other kinds of AIs (Ch. 10), Bostrom again idealizes superintelligences as God-like agents. I agree that God-like AIs will probably emerge eventually, perhaps millennia from now as a result of astroengineering. But I think they'll take time even after AI exceeds human intelligence.

Bostrom's discussion has the air of mathematical idealization more than practical engineering. For instance, he imagines that a genie AI perhaps wouldn't need to ask humans for their commands because it could simply predict them (p. 149), or that an oracle AI might be able to output the source code for a genie (p. 150). Bostrom's observations resemble crude proofs establishing the equal power of different kinds of AIs, analogous to theo-

rems about the equivalency of single-tape and multi-tape Turing machines. But Bostrom's theorizing ignores computational complexity, which would likely be immense for the kinds of God-like feats that he's imagining of his superintelligences. I don't know the computational complexity of God-like powers, but I suspect they could be bigger than Bostrom's vision implies. Along this dimension at least, I sympathize with Tom Chivers, who felt that Bostrom's book "has, in places, the air of theology: great edifices of theory built on a tiny foundation of data."

I find that I enter a different mindset when pondering pure mathematics compared with cogitating on more practical scenarios. Mathematics is closer to fiction, because you can define into existence any coherent structure and play around with it using any operation you like no matter its computational complexity. Heck, you can even, say, take the supremum of an uncountably infinite set. It can be tempting after a while to forget that these structures are mere fantasies and treat them a bit too literally. While Bostrom's gods are not obviously *only* fantasies, it would take a lot more work to argue for their realism. MIRI and FHI focus on recruiting mathematical and philosophical talent, but I think they would do well also to bring engineers into the mix, because it's all too easy to develop elaborate mathematical theories around imaginary entities.

30 How big would a superintelligence be?

To get some grounding on this question, consider a single brain emulation. Bostrom estimates that running an upload would require at least one of the fastest supercomputers by today's standards. Assume the emulation would think thousands to millions of times faster than a biological brain. Then to significantly outpace 7 billion humans (or, say, only the most educated 1 billion humans), we would need at least thousands to millions of uploads. These numbers might be a few orders of magnitude lower if the uploads are copied from a really smart person and are thinking about relevant questions with more focus than most humans. Also, Moore's law may continue to shrink computers by several orders of magnitude. Still, we might need at least the equivalent size of several of today's supercomputers to run an emulation-based AI that substantially competes with the human race.

Maybe a *de novo* AI could be significantly smaller if it's vastly more efficient than a human brain. Of course, it might also be vastly larger because it hasn't had millions of years of evolution to optimize its efficiency.

In discussing AI boxing (Ch. 9), Bostrom suggests, among other things, keeping an AI in a Faraday cage. Once the AI became superintelligent, though, this would need to be a pretty big cage.

31 Another hypothetical AI takeoff scenario

Inspired by the preceding discussion of socializing AIs, here's another scenario in which general AI follows more straightforwardly from the kind of weak AI used in Silicon Valley than in the first scenario.

- 2014: Weak AI is deployed by many technology companies for image classification, voice recognition, web search, consumer data analytics, recommending Facebook posts, personal digital assistants (PDAs), and copious other forms of automation. There's pressure to make AIs more insightful, including using deep neural networks.
- 2024: Deep learning is widespread among major tech companies. It allows for supervised learning with less manual feature engineering. Researchers develop more sophisticated forms of deep learning that can model specific kinds of systems, in-

cluding temporal dynamics. A goal is to improve generative modeling so that learning algorithms take input and not only make immediate predictions but also develop a probability distribution over what other sorts of things are happening at the same time. For instance, a Google search would not only return results but also give Google a sense of the mood, personality, and situation of the user who typed it. Of course, even in 2014, we have this in some form via Google Personalized Search, but by 2024, the modeling will be more "built in" to the learning architecture and less hand-crafted.

- 2035: PDAs using elaborate learned models are now extremely accurate at predicting what their users want. The models in these devices embody in crude form some of the same mechanisms as the user's own cognitive processes. People become more trusting of leaving their PDAs on autopilot to perform certain mundane tasks.
- 2065: A new generation of PDAs is now sufficiently sophisticated that it has a good grasp of the user's intentions. It can perform tasks as well as a human personal assistant in most cases – doing what the user wanted because it has a strong predictive model of the user's personality and goals. Meanwhile, researchers continue to unlock neural mechanisms of judgment, decision making, and value, which inform those who develop cutting-edge PDA architectures.
- 2095: PDAs are now essentially full-fledged copies of their owners. Some people have dozens of PDAs working for them, as well as meta-PDAs who help with oversight. Some PDAs make disastrous mistakes, and society debates how to construe legal accountability for PDA wrongdoing. Courts decide that owners are responsible, which makes people more cautious, but given the immense competitive pressure to outsource work to PDAs, the automation trend is not substantially affected.
- 2110: The world moves too fast for biological humans to participate. Most of the world is now run by PDAs, which – because they were built based on inferring the goals of their owners – protect their owners for the most part. However, there remains conflict among PDAs, and the world is not a completely safe place.
- 2130: PDA-led countries create a world government to forestall costly wars. The transparency of digital society allows for more credible commitments and enforcement.

I don't know what would happen with goal preservation in this scenario. Would the PDAs eventually decide to stop goal drift? Would there be any gross and irrevocable failures of translation between actual human values and what the PDAs infer? Would some people build "rogue PDAs" that operate under their own drives and that pose a threat to society? Obviously there are hundreds of ways the scenario as I described it could be varied.

32 AI: More like the economy than like robots?

What will AI look like over the next 30 years? I think it'll be similar to the Internet revolution or factory automation. Rather than developing agent-like individuals with goal systems, people will mostly optimize routine processes, developing ever more elaborate systems for mechanical tasks and information processing. The world will move very quickly – not because AI "agents" are thinking at high speeds but because software systems collectively will be capable of amazing feats. Imagine, say, bots making edits on Wikipedia that become ever more sophisticated. AI, like the economy, will be more of a network property than a localized, discrete actor.

As more and more jobs become automated, more and more people will be needed to work on the automation itself: building, maintaining, and repairing complex software and hardware systems, as well as generating training data on which to do machine learning. I expect increasing automation in software maintenance, including more robust systems and systems that detect and try to fix errors. Present-day compilers that detect syntactical problems in code offer a hint of what's possible in this regard. I also expect increasingly high-level languages and interfaces for programming computer systems. Historically we've seen this trend – from assembly language, to C, to Python, to WYSIWIG editors, to fully pre-built website styles, natural-language Google searches, and so on. Maybe eventually, as Marvin Minsky (1984) proposes, we'll have systems that can infer our wishes from high-level gestures and examples. This suggestion is redolent of my PDA scenario above.

In 100 years, there may be artificial human-like agents, and at that point more sci-fi AI images may become more relevant. But by that point the world will be very different, and I'm not sure the agents created will be discrete in the way humans are. Maybe we'll instead have a kind of global brain in which processes are much more intimately interconnected, transferable, and transparent than humans are today. Maybe there will never be a distinct AGI agent on a single supercomputer; maybe superhuman intelligence will always be a distributed system across many interacting computer systems. Robin Hanson gives an analogy in "I Still Don't Get Foom":

> Imagine in the year 1000 you didn't understand "industry," but knew it was coming, would be powerful, and involved iron and coal. You might then have pictured a blacksmith inventing and then forging himself an industry, and standing in a city square waiving it about, commanding all to bow down before his terrible weapon. Today you can see this is silly — industry sits in thousands of places, must be wielded by thousands of people, and needed thousands of inventions to make it work.
>
> Similarly, while you might imagine someday standing in awe in front of a super intelligence that embodies all the power of a new age, superintelligence just isn't the sort of thing that one project could invent. As "intelligence" is just the name we give to being better at many mental tasks by using many good mental modules, there's no one place to improve it.

Of course, this doesn't imply that humans will maintain the reins of control. Even today and throughout history, economic growth has had a life of its own. Technological development is often unstoppable even in the face of collective efforts of humanity to restrain it (e.g., nuclear weapons). In that sense, we're already familiar with humans being overpowered by forces beyond their control. An AI takeoff will represent an acceleration of this trend, but it's unclear whether the dynamic will be fundamentally discontinuous from what we've seen so far.

Gregory Stock's (1993) *Metaman*:

> While many people have had ideas about a global brain, they have tended to suppose that this can be improved or altered by humans according to their will. Metaman can be seen as a development that directs humanity's will to its own ends, whether it likes it or not, through the operation of market forces.

Vernor Vinge reported that *Metaman* helped him see how a singularity might not be completely opaque to us. Indeed, a superintelligence might look something like present-day

human society, with leaders at the top: "That apex agent itself might not appear to be much deeper than a human, but the overall organization that it is coordinating would be more creative and competent than a human." *Update, Nov. 2015*: I'm increasingly leaning toward the view that the development of AI over the coming century will be slow, incremental, and more like the Internet than like unified artificial agents. I think humans will develop vastly more powerful software tools long before highly competent autonomous agents emerge, since common-sense autonomous behavior is just so much harder to create than domain-specific tools. If this view is right, it suggests that work on AGI issues may be somewhat less important than I had thought, since

1. AGI is very far away and
2. the "unified agent" models of AGI that MIRI tends to play with might be somewhat inaccurate even once true AGI emerges.

This is a weaker form of the standard argument that "we should wait until we know more what AGI will look like to focus on the problem" and that"worrying about the dark side of artificial intelligence is like worrying about overpopulation on Mars".

I don't think the argument against focusing on AGI works because

1. some MIRI research, like on decision theory, is "timeless" (pun intended) and can be fruitfully started now
2. beginning the discussion early is important for ensuring that safety issues will be explored when the field is more mature
3. I might be wrong about slow takeoff, in which case MIRI-style work would be more important.

Still, this point does cast doubt on heuristics like "directly shaping AGI dominates all other considerations." It also means that a lot of the ways "AI safety" will play out on shorter timescales will be with issues like assassination drones, computer security, financial meltdowns, and other more mundane, catastrophic-but-not-extinction-level events.

33 Importance of whole-brain emulation

I don't currently know enough about the technological details of whole-brain emulation to competently assess predictions that have been made about its arrival dates. In general, I think prediction dates are too optimistic (planning fallacy), but it still could be that human-level emulation comes before from-scratch human-level AIs do. Of course, perhaps there would be some mix of both technologies. For instance, if crude brain emulations didn't reproduce all the functionality of actual human brains due to neglecting some cellular and molecular details, perhaps from-scratch AI techniques could help fill in the gaps.

If emulations are likely to come first, they may deserve more attention than other forms of AI. In the long run, bottom-up AI will dominate everything else, because human brains – even run at high speeds – are only so smart. But a society of brain emulations would run vastly faster than what biological humans could keep up with, so the details of shaping AI would be left up to them, and our main influence would come through shaping the emulations. Our influence on emulations could matter a lot, not only in nudging the dynamics of how emulations take off but also because the values of the emulation society might depend significantly on who was chosen to be uploaded.

One argument why emulations might improve human ability to control AI is that both emulations and the AIs they would create would be digital minds, so the emulations' AI creations wouldn't have inherent speed advantages purely due to the greater efficiency

of digital computation. Emulations' AI creations might still have more efficient mind architectures or better learning algorithms, but building those would take work. The "for free" speedup to AIs just because of their substrate would not give AIs a net advantage over emulations. Bostrom feels "This consideration is not too weighty" (p. 244 of *Superintelligence*) because emulations might still be far less intelligent than AGI. I find this claim strange, since it seems to me that the main advantage of AGI in the short run would be its speed rather than qualitative intelligence, which would take (subjective) time and effort to develop.

Bostrom also claims that if emulations come first, we would face risks from two transitions (humans to emulations, and emulations to AI) rather than one (humans to AI). There may be some validity to this, but it also seems to neglect the realization that the "AI" transition has many stages, and it's possible that emulation development would overlap with some of those stages. For instance, suppose the AI trajectory moves from $AI_1 \rightarrow AI_2 \rightarrow AI_3$. If emulations are as fast and smart as AI_1, then the transition to AI_1 is not a major risk for emulations, while it would be a big risk for humans. This is the same point as made in the previous paragraph.

"Emulation timelines and AI risk" has further discussion of the interaction between emulations and control of AIs.

34 Why work against brain-emulation risks appeals to suffering reducers

Previously in this piece I compared the expected suffering that would result from a rogue AI vs. a human-inspired AI. I suggested that while a first-guess calculation may tip in favor of a human-inspired AI on balance, this conclusion is not clear and could change with further information, especially if we had reason to think that many rogue AIs would be "minimizers" of something or would not colonize space.

In the case of brain emulations (and other highly neuromorphic AIs), we already know a lot about what those agents would look like: They would have both maximization and minimization goals, would usually want to colonize space, and might have some human-type moral sympathies (depending on their edit distance relative to a pure brain upload). The possibilities of pure-minimizer emulations or emulations that don't want to colonize space are mostly ruled out. As a result, it's pretty clear that "unsafe" brain emulations and emulation arms-race dynamics would result in more expected suffering than a more deliberative future trajectory in which altruists have a bigger influence, even if those altruists don't place particular importance on reducing suffering.

Thus, the types of interventions that pure suffering reducers would advocate with respect to brain emulations might largely match those that altruists who care about other values would advocate. This means that getting more people interested in making the brain-emulation transition safer and more humane seems like a safe bet for suffering reducers.

One might wonder whether "unsafe" brain emulations would be more likely to produce rogue AIs, but this doesn't seem to be the case, because even unfriendly brain emulations would collectively be amazingly smart and would want to preserve their own goals. Hence they would place as much emphasis on controlling their AIs as would a more human-friendly emulation world. A main exception to this is that a more cooperative, unified emulation world might be less likely to produce rogue AIs because of less pressure for arms races.

35 Would emulation work accelerate neuromorphic AI?

In Ch. 2 of *Superintelligence*, Bostrom makes a convincing case against brain-computer in-

terfaces as an easy route to significantly superhuman performance. One of his points is that it's very hard to decode neural signals in one brain and reinterpret them in software or in another brain (pp. 46-47). This might be an AI-complete problem.

But then in Ch. 11, Bostrom goes on to suggest that emulations might learn to decompose themselves into different modules that could be interfaced together (p. 172). While possible in principle, I find such a scenario implausible for the reason Bostrom outlined in Ch. 2: There would be so many neural signals to hook up to the right places, which would be different across different brains, that the task seems hopelessly complicated to me. Much easier to build something from scratch.

Along the same lines, I doubt that brain emulation in itself would vastly accelerate neuromorphic AI, because emulation work is mostly about copying without insight. *Cognitive psychology* is often more informative about AI architectures than cellular neuroscience, because general psychological systems can be understood in functional terms as inspiration for AI designs, compared with the opacity of neuronal spaghetti. In Bostrom's list of examples of AI techniques inspired by biology (Ch. 14, "Technology couplings"), only a few came from neuroscience specifically. That said, emulation work might involve some cross-pollination with AI, and in any case, it might accelerate interest in brain/artificial intelligence more generally or might put pressure on AI groups to move ahead faster. Or it could funnel resources and scientists away from *de novo* AI work. The upshot isn't obvious.

A "Singularity Summit 2011 Workshop Report" includes the argument that neuromorphic AI should be easier than brain emulation because "Merely reverse-engineering the Microsoft Windows code base is hard, so reverse-engineering the brain is probably much harder" (Salamon & Muehlhauser, 2012). But emulation is not reverse-engineering. As Robin Hanson (1994) explains, brain emulation is more akin to porting software (though probably "emulation" actually is the more precise word, since emulation involves simulating the original hardware). While I don't know any fully reverse-engineered versions of Windows, there are several Windows emulators, such as VirtualBox.

Of course, if emulations emerged, their significantly faster rates of thinking would multiply progress on non-emulation AGI by orders of magnitude. Getting safe emulations doesn't by itself get safe *de novo* AGI because the problem is just pushed a step back, but we could leave AGI work up to the vastly faster emulations. Thus, for biological humans, if emulations come first, then influencing their development is the last thing we ever need to do. That said, thinking several steps ahead about what kinds of AGIs emulations are likely to produce is an essential part of influencing emulation development in better directions.

36 Are neuromorphic or mathematical AIs more controllable?

Arguments for mathematical AIs:

- Behavior and goals are more transparent, and goal preservation seems easier to specify (see "The Ethics of Artificial Intelligence" by Bostrom and Yudkowsky, p. 16).
- Neuromorphic AIs might speed up mathematical AI, leaving less time to figure out control.

Arguments for neuromorphic AIs:

- We understand human psychology, expectations, norms, and patterns of behavior. Mathematical AIs could be totally alien and hence unpredictable.
- If neuromorphic AIs came first, they could think faster and help figure out goal preservation, which I assume does require mathematical AIs at the end of the day.
- Mathematical AIs may be more prone to

unexpected breakthroughs that yield radical jumps in intelligence.

In the limit of very human-like neuromorphic AIs, we face similar considerations as between emulations vs. from-scratch AIs – a tradeoff which is not at all obvious.

Overall, I think mathematical AI has a better best case but also a worse worst case than neuromorphic. If you really want goal preservation and think goal drift would make the future worthless, you might lean more towards mathematical AI because it's more likely to perfect goal preservation. But I probably care less about goal preservation and more about avoiding terrible outcomes.

In *Superintelligence* (Ch. 14), Bostrom comes down strongly in favor of mathematical AI being safer. I'm puzzled by his high degree of confidence here. Bostrom claims that unlike emulations, neuromorphic AIs wouldn't have human motivations by default. But this seems to depend on how human motivations are encoded and what parts of human brains are modeled in the AIs.

In contrast to Bostrom, a 2011 Singularity Summit workshop ranked neuromorphic AI as more controllable than (non-friendly) mathematical AI, though of course they found friendly mathematical AI most controllable (Salamon & Muehlhauser, 2012). The workshop's aggregated probability of a good outcome given brain emulation or neuromorphic AI turned out to be the same (14%) as that for mathematical AI (which might be either friendly or unfriendly).

37 Impacts of empathy for AIs

As I noted above, advanced AIs will be complex agents with their own goals and values, and these will matter ethically. Parallel to discussions of robot rebellion in science fiction are discussions of robot rights. I think even present-day computers deserve a tiny bit of moral concern, and complex computers of the future will command even more ethical consideration.

How might ethical concern for machines interact with control measures for machines?

37.1 Slower AGI development?

As more people grant moral status to AIs, there will likely be more scrutiny of AI research, analogous to how animal activists in the present monitor animal testing. This may make AI research slightly more difficult and may distort what kinds of AIs are built depending on the degree of empathy people have for different types of AIs (Calverley, 2005). For instance, if few people care about invisible, non-embodied systems, researchers who build these will face less opposition than those who pioneer suffering robots or animated characters that arouse greater empathy. If this possibility materializes, it would contradict present trends where it's often helpful to create at least a toy robot or animated interface in order to "sell" your research to grant-makers and the public.

Since it seems likely that reducing the pace of progress toward AGI is on balance beneficial, a slowdown due to ethical constraints may be welcome. Of course, depending on the details, the effect could be harmful. For instance, perhaps China wouldn't have many ethical constraints, so ethical restrictions in the West might slightly favor AGI development by China and other less democratic countries. (This is not guaranteed. For what it's worth, China has already made strides toward reducing animal testing.)

In any case, I expect ethical restrictions on AI development to be small or nonexistent until many decades from now when AIs develop perhaps mammal-level intelligence. So maybe such restrictions won't have a big impact on AGI progress. Moreover, it may be that most AGIs will be sufficiently alien that they won't arouse much human sympathy.

Brain emulations seem more likely to raise

ethical debate because it's much easier to argue for their personhood. If we think brain emulation coming before AGI is good, a slowdown of emulations could be unfortunate, while if we want AGI to come first, a slowdown of emulations should be encouraged.

Of course, emulations and AGIs do actually matter and deserve rights in principle. Moreover, movements to extend rights to machines in the near term may have long-term impacts on how much post-humans care about suffering subroutines run at galactic scale. I'm just pointing out here that ethical concern for AGIs and emulations also may somewhat affect timing of these technologies.

37.2 Attitudes toward AGI control

Most humans have no qualms about shutting down and rewriting programs that don't work as intended, but many do strongly object to killing people with disabilities and designing better-performing babies. Where to draw a line between these cases is a tough question, but as AGIs become more animal-like, there may be increasing moral outrage at shutting them down and tinkering with them willy nilly.

Nikola Danaylov asked Roman Yampolskiy whether it was speciesist or discrimination in favor of biological beings to lock up machines and observe them to ensure their safety before letting them loose.

At a lecture in Berkeley, CA, Nick Bostrom was asked whether it's unethical to "chain" AIs by forcing them to have the values we want. Bostrom replied that we have to give machines *some* values, so they may as well align with ours. I suspect most people would agree with this, but the question becomes trickier when we consider turning off erroneous AGIs that we've already created because they don't behave how we want them to. A few hard-core AGI-rights advocates might raise concerns here. More generally, there's a segment of transhumanists (including young Eliezer Yudkowsky) who feel that human concerns are overly parochial and that it's chuavanist to impose our "monkey dreams" on an AGI, which is the next stage of evolution.

The question is similar to whether one sympathizes with the Native Americans (humans) or their European conquerors (rogue AGIs). Before the second half of the 20th century, many history books glorified the winners (Europeans). After a brief period in which humans are quashed by a rogue AGI, its own "history books" will celebrate its conquest and the bending of the arc of history toward "higher", "better" forms of intelligence. (In practice, the psychology of a rogue AGI probably wouldn't be sufficiently similar to human psychology for these statements to apply literally, but they would be true in a metaphorical and implicit sense.)

David Althaus worries that if people sympathize too much with machines, society will be less afraid of an AI takeover, even if AI takeover is bad on purely altruistic grounds. I'm less concerned about this because even if people agree that advanced machines are sentient, they would still find it intolerable for AGIs to commit speciecide against humanity. Everyone agrees that Hitler was sentient, after all. Also, if it turns out that rogue-AI takeover is altruistically desirable, it would be better if more people agreed with this, though I expect an extremely tiny fraction of the population would ever come around to such a position.

Where sympathy for AGIs might have more impact is in cases of softer takeoff where AGIs work in the human economy and acquire increasing shares of wealth. The more humans care about AGIs for their own sakes, the more such transitions might be tolerated. Or would they? Maybe seeing AGIs as more human-like would evoke the xenophobia and ethnic hatred that we've seen throughout history whenever a group of people gains wealth (e.g., Jews in Medieval Europe) or steals jobs (e.g., immigrants of various types throughout history).

Personally, I think greater sympathy for AGI

is likely net positive because it may help allay anti-alien prejudices that may make cooperation with AGIs harder. When a *Homo sapiens* tribe confronts an outgroup, often it reacts violently in an effort to destroy the evil foreigners. If instead humans could cooperate with their emerging AGI brethren, better outcomes would likely follow.

38 Charities working on this issue

What are some places where donors can contribute to make a difference on AI? The Center on Long-Term Risk (CLR) explores questions like these, though at the moment the organization is rather small. MIRI is larger and has a longer track record. Its values are more conventional, but it recognizes the importance of positive-sum opportunities to help many values systems, which includes suffering reduction. More reflection on these topics can potentially reduce suffering and further goals like eudaimonia, fun, and interesting complexity at the same time.

Because AI is affected by many sectors of society, these problems can be tackled from diverse angles. Many groups besides FRI and MIRI examine important topics as well, and these organizations should be explored further as potential charity recommendations.

39 Is MIRI's work too theoretical?

Most of MIRI's publications since roughly 2012 have focused on formal mathematics, such as logic and provability. These are tools not normally used in AGI research. I think MIRI's motivations for this theoretical focus are

1. Pessimism about the problem difficulty: Luke Muehlhauser writes that "Especially for something as complex as Friendly AI, our message is: 'If we prove it correct, it *might* work. If we *don't* prove it correct, it *definitely* won't work.'"
2. Not speeding unsafe AGI: Building real-world systems would contribute toward non-safe AGI research.
3. Long-term focus: MIRI doesn't just want a system that's the next level better but aims to explore the theoretical limits of possibilities.

I personally think reason #3 is most compelling. I doubt #2 is hugely important given MIRI's small size, though it matters to some degree. #1 seems a reasonable strategy in moderation, though I favor approaches that look decently likely to yield non-terrible outcomes rather than shooting for the absolute best outcomes.

Software can be proved correct, and sometimes this is done for mission-critical components, but most software is not validated. I suspect that AGI will be sufficiently big and complicated that proving safety will be impossible for humans to do completely, though I don't rule out the possibility of software that would help with correctness proofs on large systems. Muehlhauser and comments on his post largely agree with this.

What kind of track record does theoretical mathematical research have for practical impact? There are certainly several domains that come to mind, such as the following.

- Auction game theory has made governments billions of dollars and is widely used in Internet advertising.
- Theoretical physics has led to numerous forms of technology, including electricity, lasers, and atomic bombs. However, immediate technological implications of the most theoretical forms of physics (string theory, Higgs boson, black holes, etc.) are less pronounced.
- Formalizations of many areas of computer science have helped guide practical implementations, such as in algorithm complexity, concurrency, distributed systems, cryptography, hardware verification, and

> so on. That said, there are also areas of theoretical computer science that have little immediate application. Most software engineers only know a little bit about more abstract theory and still do fine building systems, although if no one knew theory well enough to design theory-based tools, the software field would be in considerably worse shape.

All told, I think it's important for someone to do the kinds of investigation that MIRI is undertaking. I personally would probably invest more resources than MIRI is in hacky, approximate solutions to AGI safety that don't make such strong assumptions about the theoretical cleanliness and soundness of the agents in question. But I expect this kind of less perfectionist work on AGI control will increase as more people become interested in AGI safety.

There does seem to be a significant divide between the math-oriented conception of AGI and the engineering/neuroscience conception. Ben Goertzel takes the latter stance:

> I strongly suspect that to achieve high levels of general intelligence using realistically limited computational resources, one is going to need to build systems with a nontrivial degree offundamental unpredictabilityto them. This is what neuroscience suggests, it's what my concrete AGI design work suggests, and it's what my theoretical work on GOLEM and related ideas suggests (Goertzel, 2014). And none of the public output of SIAI researchers or enthusiasts has given me any reason to believe otherwise, yet.

Personally I think Goertzel is more likely to be right on this particular question. Those who view AGI as fundamentally complex have more concrete results to show, and their approach is by far more mainstream among computer scientists and neuroscientists. Of course, proofs about theoretical models like Turing machines and lambda calculus are also mainstream, and few can dispute their importance. But Turing-machine theorems do little to constrain our understanding of what AGI will actually look like in the next few centuries. That said, there's significant peer disagreement on this topic, so epistemic modesty is warranted. In addition, *if* the MIRI view is right, we might have more scope to make an impact to AGI safety, and it would be possible that important discoveries could result from a few mathematical insights rather than lots of detailed engineering work. Also, most AGI research is more engineering-oriented, so MIRI's distinctive focus on theory, especially abstract topics like decision theory, may target an underfunded portion of the space of AGI-safety research.

In "How to Study Unsafe AGI's safely (and why we might have no choice)", Punoxysm makes several points that I agree with, including that AGI research is likely to yield many false starts before something self-sustaining takes off, and those false starts could afford us the opportunity to learn about AGI experimentally. Moreover, this kind of ad-hoc, empirical work may be necessary if, as seems to me probable, fully rigorous mathematical models of safety aren't sufficiently advanced by the time AGI arrives.

Ben Goertzel likewise suggests that a fruitful way to approach AGI control is to study small systems and "in the usual manner of science, attempt to arrive at a solid theory of AGI intelligence and ethics based on a combination of conceptual and experimental-data considerations". He considers this view the norm among "most AI researchers or futurists". I think empirical investigation of how AGIs behave is very useful, but we also have to remember that many AI scientists are overly biased toward "build first; ask questions later" because

- building may be more fun and exciting

than worrying about safety (Steven M. Bellovin observed with reference to open-source projects: "Quality takes work, design, review and testing and those are not nearly as much fun as coding".)
- there's more incentive from commercial applications and government grants to build rather than introspect
- scientists may want AGI sooner so that they personally or their children can reap its benefits.

On a personal level, I suggest that if you really like building systems rather than thinking about safety, you might do well to earn to give in software and donate toward AGI-safety organizations.

40 Next steps

Here are some rough suggestions for how I recommend proceeding on AGI issues and, in [brackets], roughly how long I expect each stage to take. Of course, the stages needn't be done in a strict serial order, and step 1 should continue indefinitely, as we continue learning more about AGI from subsequent steps.

1. *Decide if we want human-controlled, goal-preserving AGI [5-10 years].* This involves exploring questions about what types of AGI scenarios might unfold and how much suffering would result from AGIs of various types.
2. *Assuming we decide we do want controlled AGI: Network with academics and AGI developers to raise the topic and canvass ideas [5-10 years].* We could reach out to academic AGI-like projects, including these listed by Pei Wang and these listed on Wikipedia, as well as to machine ethics and roboethics communities. There are already some discussions about safety issues among these groups, but I would expand the dialogue, have private conversations, write publications, hold conferences, etc. These efforts both inform us about the lay of the land and build connections in a friendly, mutualistic way.
3. *Lobby for greater funding of research into AGI safety [10-20 years].* Once the idea and field of AGI safety have become more mainstream, it should be possible to differentially speed up safety research by getting more funding for it – both from governments and philanthropists. This is already somewhat feasible; for instance: "In 2014, the US Office of Naval Research announced that it would distribute $7.5 million in grants over five years to university researchers to study questions of machine ethics as applied to autonomous robots."
4. *The movement snowballs [decades].* It's hard to plan this far ahead, but I imagine that eventually (within 25-50 years?) AGI safety will become a mainstream political topic in a similar way as nuclear security is today. Governments may take over in driving the work, perhaps with heavy involvement from companies like Google. This is just a prediction, and the actual way things unfold could be different.

I recommend avoiding a confrontational approach with AGI developers. I would not try to lobby for restrictions on their research (in the short term at least), nor try to "slow them down" in other ways. AGI developers are the allies we need most at this stage, and most of them don't want uncontrolled AGI either. Typically they just don't see their work as risky, and I agree that at this point, no AGI project looks set to unveil something dangerous in the next decade or two. For many researchers, AGI is a dream they can't help but pursue. Hopefully we can engender a similar enthusiasm about pursuing AGI safety.

In the longer term, tides may change, and perhaps many AGI developers will desire

government-imposed restrictions as their technologies become increasingly powerful. Even then, I'm doubtful that governments will be able to completely control AGI development (see, e.g., the criticisms by John McGinnis of this approach), so differentially pushing for more safety work may continue to be the most leveraged solution. History provides a poor track record of governments refraining from developing technologies due to ethical concerns; (Eckersley & Sandberg, 2014, p. 187) (p. 187) cite "human cloning and land-based autonomous robotic weapons" as two of the few exceptions, with neither prohibition having a long track record.

I think the main way in which we should try to affect the speed of regular AGI work is by aiming to avoid setting off an AGI arms race, either via an AGI Sputnik moment or else by more gradual diffusion of alarm among world militaries. It's possible that discussing AGI scenarios too much with military leaders could exacerbate a militarized reaction. If militaries set their sights on AGI the way the US and Soviet Union did on the space race or nuclear-arms race during the Cold War, the amount of funding for unsafe AGI research might multiply by a factor of 10 or maybe 100, and it would be aimed in harmful directions.

41 Where to push for maximal impact?

Here are some candidates for the best object-level projects that altruists could work on with reference to AI. Because AI seems so crucial, these are also candidates for the best object-level projects in general. Meta-level projects like movement-building, career advice, earning to give, fundraising, etc. are also competitive. I've scored each project area out of 10 points to express a rough subjective guess of the value of the work for suffering reducers.

Research whether controlled or uncontrolled AI yields more suffering (score = 10/10)

Pros:

- Figuring out which outcome is better should come before pushing ahead too far in any particular direction.
- This question remains non-obvious and so has very high expected value of information.
- None of the existing big names in AI safety have explored this question because reducing suffering is not the dominant priority for them.

Cons:

- None.

Push for suffering-focused AI-safety approaches (score = 10/10)

Most discussions of AI safety assume that human extinction and failure to spread (human-type) eudaimonia are the main costs of takeover by uncontrolled AI. But as noted in this piece, AIs would also spread astronomical amounts of suffering. Currently no organization besides FRI is focused on how to do AI safety work with the primary aim of avoiding outcomes containing huge amounts of suffering.

One example of a suffering-focused AI-safety approach is to design AIs so that even if they do get out of control, they "fail safe" in the sense of not spreading massive amounts of suffering into the cosmos. For example:

1. AIs should be inhibited from colonizing space, or if they do colonize space, they should do so in less harmful ways.
2. "Minimizer" utility functions have less risk of creating new universes than "maximizer" ones do.
3. Simpler utility functions (e.g., creating uniform paperclips) might require fewer

suffering subroutines than complex utility functions would.

4. AIs with expensive intrinsic values (e.g., maximize paperclips) may run fewer complex minds than AIs with cheaper values (e.g., create at least one paperclip on each planet), because AIs with cheaper values have lower opportunity cost for using resources and so can expend more of their cosmic endowment on learning about the universe to make sure they've accomplished their goals properly. (Thanks to a friend for this point.) From this standpoint, suffering reducers might prefer an AI that aims to "maximize paperclips" over one that aims to "make sure there's at least one paperclip per planet." However, perhaps the paperclip maximizer would prefer to create new universes, while the "at least one paperclip per planet" AI wouldn't; indeed, the "one paperclip per planet" AI might prefer to have a smaller multiverse so that there would be fewer planets that don't contain paperclips. Also, the satisficing AI would be easier to compromise with than the maximizing AI, since the satisficer's goals could be carried out more cheaply. There are other possibilities to consider as well. Maybe an AI with the instructions to "be 70% sure of having made one paperclip and then shut down all of your space-colonization plans" would not create much suffering (depending on how scrupulous the AI was about making sure that what it had created was really a paperclip, that it understood physics properly, etc.).

The problem with bullet #1 is that *if* you can succeed in preventing AGIs from colonizing space, it seems like you should already have been able to control the AGI altogether, since the two problems appear about equally hard. But maybe there are clever ideas we haven't thought of for reducing the spread of suffering even if humans lose total control.

Another challenge is that those who don't place priority on reducing suffering may not agree with these proposals. For example, I would guess that most AI scientists would say, "If the AGI kills humans, at least we should ensure that it spreads life into space, creates a complex array of intricate structures, and increases the size of our multiverse."

Work on AI control and value-loading problems (score = 4/10)

Pros:

- At present, controlled AI seems more likely good than bad.
- Relatively little work thus far, so marginal effort may make a big impact.

Cons:

- It may turn out that AI control increases net expected suffering.
- This topic may become a massive area of investment in coming decades, because everyone should theoretically care about it. Maybe there's more leverage in pushing on neglected areas of particular concern for suffering reduction.

Research technological/economic/ political dynamics of an AI takeoff and push in better directions (score = 3/10)

By this I have in mind scenarios like those of Robin Hanson for emulation takeoff, or Bostrom's (2004) "The Future of Human Evolution".

Pros:

- Many scenarios have not been mapped out. There's a need to introduce economic/social realism to AI scenarios, which at present often focus on technical challenges and idealized systems.

- Potential to steer dynamics in more win-win directions.

Cons:

- Broad subject area. Work may be somewhat replaceable as other researchers get on board in the coming decades.
- More people have their eyes on general economic/social trends than on specific AI technicalities, so there may be lower marginal returns to additional work in this area.
- While technological progress is probably the biggest influence on history, it's also one of the more inevitable influences, making it unclear how much we can affect it. Our main impact on it would seem to come through differential technological progress. In contrast, values, institutions, and social movements can go in many different directions depending on our choices.

Promote the ideal of cooperation on AI values (e.g., CEV by Yudkowsky (2004)) (score = 2/10)

Pros:

- Whereas technical work on AI safety is of interest to and benefits everyone – including militaries and companies with non-altruistic aims – promoting CEV is more important to altruists. I don't see CEV as a likely outcome even if AI is controlled, because it's more plausible that individuals and groups will push for their own agendas.

Cons:

- It's very hard to achieve CEV. It depends on a lot of really complex political and economic dynamics that millions of altruists are already working to improve.
- Promoting CEV as an ideal to approximate may be confused in people's minds with suggesting that CEV is likely to happen. The latter assumption is probably wrong and so may distort people's beliefs about other crucial questions. For instance, if CEV was likely, then it would be more likely that suffering reducers should favor controlled AI; but the fact of the matter is that anything more than crude approximations to CEV will probably not happen.

Promote a smoother, safer takeoff for brain emulation (score = 2/10)

Pros:

- As noted above, it's more plausible that suffering reducers should favor emulation safety than AI safety.
- The topic seems less explored than safety of *de novo* AIs.

Cons:

- I find it slightly more likely that *de novo* AI will come first, in which case this work wouldn't be as relevant. In addition, AI may have more impacts on society even before it reaches the human level, again making it slightly more relevant.
- Safety measures might require more political and less technical work, in which case it's more likely to be done correctly by policy makers in due time. The value-loading problem seems much easier for emulations because it might just work to upload people with good values, assuming no major value corruption during or after uploading.
- Emulation is more dependent on relatively straightforward engineering improvements and less on unpredictable insight than AI. Thus, it has a clearer development timeline, so there's less urgency to investigate issues ahead of time to prepare

for an unexpected breakthrough.

Influence the moral values of those likely to control AI (score = 2/10)

Pros:

- Altruists, and especially those with niche values, may want to push AI development in more compassionate directions. This could make sense because altruists are most interested in ethics, while even power-hungry states and money-hungry individuals should care about AI safety in the long run.

Cons:

- This strategy is less cooperative. It's akin to defecting in a tragedy of the commons – pushing more for what you want rather than what everyone wants. If you do push for what everyone wants, then I would consider such work more like the "Promote the ideal of cooperation" item.
- Empirically, there isn't enough investment in other fundamental AI issues, and those may be more important than further engaging already well trodden ethical debates.

Promote a singleton over multipolar dynamics (score = 1/10)

Pros:

- A singleton, whether controlled or uncontrolled, would reduce the risk of conflicts that cause cosmic damage.

Unclear:

- There are many ways to promote a singleton. Encouraging cooperation on AI development would improve pluralism and human control in the outcome. Faster development by the leading AI project might also increase the chance of a singleton while reducing the probability of human control of the outcome. Stronger government regulation, surveillance, and coordination would increase chances of a singleton, as would global cooperation.

Cons:

- Speeding up the leading AI project might exacerbate AI arms races. And in any event, it's currently far too early to predict what group will lead the AI race.

Other variations

In general, there are several levers that we can pull on:

- safety
- arrival time relative to other technologies
- influencing values
- cooperation
- shaping social dynamics
- raising awareness
- etc.

These can be applied to any of

- *de novo* AI
- brain emulation
- other key technologies
- etc.

42 Is it valuable to work at or influence an AGI company?

Projects like DeepMind, Vicarious, OpenCog, and the AGI research teams at Google, Facebook, etc. are some of the leaders in AGI technology. Sometimes it's proposed that since these teams *might* ultimately develop AGI, altruists should consider working for, or at least lobbying, these companies so that they think more about AGI safety.

One's assessment of this proposal depends on one's view about AGI takeoff. My own opinion may be somewhat in the minority relative to expert surveys (Müller & Bostrom, 2016), but I'd be surprised if we had human-level AGI before 50 years from now, and my

median estimate might be like $\sim$ 90 years from now. That said, the idea of AGI arriving at a single point in time is probably a wrong framing of the question. Already machines are super-human in some domains, while their abilities are far below humans' in other domains. Over the coming decades, we'll see lots of advancement in machine capabilities in various fields at various speeds, without any *single point* where machines suddenly develop human-level abilities across all domains. Gradual AI progress over the coming decades will radically transform society, resulting in many small "intelligence explosions" in various specific areas, long before machines completely surpass humans overall.

In light of my picture of AGI, I think of DeepMind, Vicarious, etc. as ripples in a long-term wave of increasing machine capabilities. It seems extremely unlikely that any one of these companies or its AGI system will bootstrap itself to world dominance on its own. Therefore, I think influencing these companies with an eye toward "shaping the AGI that will take over the world" is probably naive. That said, insofar as these companies will influence the long-term trajectory of AGI research, and insofar as people at these companies are important players in the AGI community, I think influencing them has value – just not vastly more value than influencing other powerful people.

That said, as noted previously, early work on AGI safety has the biggest payoff in scenarios where AGI takes off earlier and harder than people expected. If the marginal returns to additional safety research are many times higher in these "early AGI" scenarios, then it could still make sense to put some investment into them even if they seem very unlikely.

43 Should suffering reducers focus on AGI safety?

If, upon further analysis, it looks like AGI safety would increase expected suffering, then the answer would be clear: Suffering reducers shouldn't contribute toward AGI safety and should worry somewhat about how their messages might incline others in that direction. However, I find it reasonably likely that suffering reducers will conclude that the benefits of AGI safety outweigh the risks. In that case, they would face a question of whether to push on AGI safety or on other projects that also seem valuable.

Reasons to focus on other projects:

- There are several really smart people working on AGI safety right now. The number of brilliant altruists focused on AGI safety probably exceeds the number of brilliant altruists focused on reducing suffering in the far future by several times over. Thus, it seems plausible that there remain more low-hanging fruit for suffering reducers to focus on other crucial considerations rather than delving into the technical details of implementing AGI safety.
- I expect that AGI safety will require at least, say, thousands of researchers and hundreds of thousands of programmers to get right. AGI safety is a much harder problem than ordinary computer security, and computer security demand is already very high: "In 2012, there were more than 67,400 separate postings for cybersecurity-related jobs in a range of industries". Of course, that AGI safety will need tons of researchers eventually needn't discount the value of early work, and indeed, someone who helps grow the movement to a large size would contribute as much as many detail-oriented AGI safety researchers later.

Reasons to focus on AGI safety:

- Most other major problems are also already being tackled by lots of smart people.
- AGI safety is a cause that many value systems can get behind, so working on it can be seen as more "nice" than focusing on areas that are more specific to suffering-reduction values.

All told, I would probably pursue a mixed strategy: Work primarily on questions specific to suffering reduction, but direct donations and resources toward AGI safety when opportunities arise. Some suffering reducers particularly suited to work on AGI safety could go in that direction while others continue searching for points of leverage not specific to controlling AGI.

44 Acknowledgments

Parts of this piece were inspired by discussions with various people, including David Althaus, Daniel Dewey, and Caspar Oesterheld.

References

Armstrong, S., Soares, N., Fallenstein, B., & Yudkowsky, E. (2015). Corrigibility. In *AAAI Publications.* Austin, TX, USA.

Bloom, P. (2013). *Just babies: The origins of good and evil.* New York: Crown.

Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, *15*(03), 308–314.

Bostrom, N. (2004). The future of human evolution. In C. Tandy (Ed.), *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing* (pp. 339–371). Palo Alto, California: Ria University Press.

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

Bostrom, N. (2010). *Anthropic bias: Observation selection effects in science and philosophy* (1edition ed.). Abingdon, Oxon: Routledge.

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

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316–334). Cambridge University Press.

Brooks, F. P., Jr. (1995). *The Mythical Man-month (Anniversary Ed.).* Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Calverley, D. J. (2005). Android science and the animal rights movement: are there analogies. In *Cognitive sciences society workshop, Stresa, Italy* (pp. 127–136).

Davis, E. (2015). Ethical guidelines for a superintelligence. *Artificial Intelligence*, *220*, 121–124.

Dennett, D. C. (1992). *Consciousness Explained* (1st ed.). Boston: Back Bay Books.

Eckersley, P., & Sandberg, A. (2014). Is Brain Emulation Dangerous? *Journal of Artificial General Intelligence*, *4*(3), 170–194.

Goertzel, B. (2014). GOLEM: towards an AGI meta-architecture enabling both goal preservation and radical self-improvement. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 391–403.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. *Advances in computers*, *6*, 31–88.

Good, I. J. (1982). Ethical machines. In *Tenth Machine Intelligence Workshop, Cleveland, Ohio* (Vol. 246, pp. 555–560).

Hall, J. S. (2008). Engineering utopia. *Frontiers in Artificial Intelligence and Applications*, *171*, 460.

Hanson, R. (1994). If uploads come first. *Ex-*

tropy, *6*(2), 10–15.

Hanson, R., & Yudkowsky, E. (2013). The Hanson-Yudkowsky AI-Foom debate. *Berkeley, CA: Machine Intelligence Research Institute*.

Kaplan, R. D. (2013). *The revenge of geography: What the map tells us about coming conflicts and the battle against fate* (Reprint edition ed.). Random House Trade Paperbacks.

Kurzweil, R. (2000). *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. New York: Penguin Books.

Minsky, M. (1984). Afterword to vernor vinge's novel, "True Names.". *Unpublished manuscript*.

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 553–571). Berlin: Springer.

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In (pp. 663–670).

Russell, S. J., Norvig, P., Canny, J. F., Malik, J. M., & Edwards, D. D. (2003). *Artificial intelligence: a modern approach* (Vol. 2). Prentice hall Upper Saddle River.

Salamon, A., & Muehlhauser, L. (2012). *Singularity summit 2011 workshop report* (Technical Report No. 1). San Francisco, CA: The Singularity Institute.

Sotala, K. (2012). Advantages of artificial intelligences, uploads, and digital minds. *International Journal of Machine Consciousness*, *04*(01), 275–291. doi: 10.1142/S1793843012400161

Stock, G. (1993). *Metaman: the merging of humans and machines into a global superorganism*. New York: Simon & Schuster.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Winnefeld, J. A., & Kendall, F. (2013). Unmanned systems integrated roadmap FY 2013-2036. *Office of the Secretary of Defense, US*.

Yudkowsky, E. (2004). Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*.

Yudkowsky, E. (2011). Complex value systems in friendly AI. In D. Hutchison et al. (Eds.), *Artificial General Intelligence* (Vol. 6830, pp. 388–393). Springer Berlin Heidelberg.

Yudkowsky, E. (2013). Intelligence explosion microeconomics. *Machine Intelligence Research Institute, accessed online October*, *23*, 2015.