
# Intelligence Explosion Microeconomics

Eliezer Yudkowsky
*Machine Intelligence Research Institute*

## Abstract

I. J. Good's thesis of the "intelligence explosion" states that a sufficiently advanced machine intelligence could build a smarter version of itself, which could in turn build an even smarter version, and that this process could continue to the point of vastly exceeding human intelligence. As Sandberg (2010) correctly notes, there have been several attempts to lay down return on investment formulas intended to represent sharp speedups in economic or technological growth, but very little attempt has been made to deal formally with Good's intelligence explosion thesis as such.

I identify the key issue as *returns on cognitive reinvestment*—the ability to invest more computing power, faster computers, or improved cognitive algorithms to yield cognitive labor which produces larger brains, faster brains, or better mind designs. There are many phenomena in the world which have been argued to be evidentially relevant to this question, from the observed course of hominid evolution, to Moore's Law, to the competence over time of machine chess-playing systems, and many more. I go into some depth on some debates which then arise on how to interpret such evidence. I propose that the next step in analyzing positions on the intelligence explosion would be to formalize return on investment curves, so that each stance can formally state which possible microfoundations they hold to be *falsified* by historical observations. More generally,

Yudkowsky, Eliezer. 2013. *Intelligence Explosion Microeconomics*. Technical report 2013-1. Berkeley, CA: Machine Intelligence Research Institute. Last modified September 13, 2013.

I pose multiple open questions of "returns on cognitive reinvestment" or "intelligence explosion microeconomics." Although such questions have received little attention thus far, they seem highly relevant to policy choices affecting outcomes for Earth-originating intelligent life.

## Contents

# 1. The Intelligence Explosion: Growth Rates of Cognitive Reinvestment

In 1965, I. J. Good[1] published a paper titled "Speculations Concerning the First Ultraintelligent Machine" (Good 1965) containing the paragraph:

> 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; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.

Many have since gone on to question Good's unquestionable, and the state of the debate has developed considerably since 1965. While waiting on Nick Bostrom's forthcoming book on the intelligence explosion, I would meanwhile recommend the survey paper "Intelligence Explosion: Evidence and Import" (Muehlhauser and Salamon 2012) for a compact overview. See also David Chalmers's (2010) paper, the responses, and Chalmers's (2012) reply.

Please note that the intelligence explosion is *not* the same thesis as a general economic or technological speedup, which is now often termed a "Singularity." Economic speedups arise in many models of the future, some of them already well formalized. For example, Robin Hanson's (1998a) "Economic Growth Given Machine Intelligence" considers emulations of scanned human brains (a.k.a. *ems*): Hanson proposes equations to model the behavior of an economy when capital (computers) can be freely converted into human-equivalent skilled labor (by running em software). Hanson concludes that the result should be a global economy with a doubling time on the order of months. This may sound startling already, but Hanson's paper doesn't try to model an agent that is *smarter* than any existing human, or whether that agent would be able to invent still-smarter agents.

The question of what happens when smarter-than-human agencies[2] are driving scientific and technological progress is difficult enough that previous attempts at formal

1. Isadore Jacob Gudak, who anglicized his name to Irving John Good and used I. J. Good for publication. He was among the first advocates of the Bayesian approach to statistics, and worked with Alan Turing on early computer designs. Within computer science his name is immortalized in the Good-Turing frequency estimator.

2. I use the term "agency" rather than "agent" to include well-coordinated groups of agents, rather than assuming a singular intelligence.

futurological modeling have entirely ignored it, although it is often discussed informally; likewise, the prospect of smarter agencies producing even smarter agencies has not been formally modeled. In his paper overviewing formal and semiformal models of technological speedup, Sandberg (2010) concludes:

> There is a notable lack of models of how an intelligence explosion could occur. This might be the most important and hardest problem to crack. . . . Most important since the emergence of superintelligence has the greatest potential of being fundamentally game-changing for humanity (for good or ill). Hardest, since it appears to require an understanding of the general nature of super-human minds or at least a way to bound their capacities and growth rates.

For responses to some arguments that the intelligence explosion is *qualitatively* forbidden—for example, because of Gödel's Theorem prohibiting the construction of artificial minds[3]—see again Chalmers (2010) or Muehlhauser and Salamon (2012). The Open Problem posed here is the *quantitative* issue: whether it's possible to get sustained returns on reinvesting cognitive improvements into further improving cognition. As Chalmers (2012) put it:

> The key issue is the "proportionality thesis" saying that among systems of certain class, an increase of $\delta$ in intelligence will yield an increase of $\delta$ in the intelligence of systems that these systems can design.

To illustrate the core question, let us consider a nuclear pile undergoing a fission reaction.[4] The first human-made critical fission reaction took place on December 2, 1942, in a rackets court at the University of Chicago, in a giant doorknob-shaped pile of uranium bricks and graphite bricks. The key number for the pile was the effective neutron multiplication factor $k$—the average number of neutrons emitted by the average number of fissions caused by one neutron. (One might consider $k$ to be the "return on investment" for neutrons.) A pile with $k > 1$ would be "critical" and increase exponentially in neutrons. Adding more uranium bricks increased $k$, since it gave a neutron more opportunity to strike more uranium atoms before exiting the pile.

Fermi had calculated that the pile ought to go critical between layers 56 and 57 of uranium bricks, but as layer 57 was added, wooden rods covered with neutron-absorbing

3. A.k.a. general AI, a.k.a. strong AI, a.k.a. Artificial General Intelligence. See Pennachin and Goertzel (2007).

4. Uranium atoms are not intelligent, so this is not meant to imply that an intelligence explosion ought to be similar to a nuclear pile. No argument by analogy is intended—just to start with a simple process on the way to a more complicated one.

cadmium foil were inserted to prevent the pile from becoming critical. The actual critical reaction occurred as the result of slowly pulling out a neutron-absorbing rod in six-inch intervals. As the rod was successively pulled out and $k$ increased, the overall neutron level of the pile increased, then leveled off each time to a new steady state. At 3:25 p.m., Fermi ordered the rod pulled out another twelve inches, remarking, "Now it will become self-sustaining. The trace will climb and continue to climb. It will not level off" (Rhodes 1986). This prediction was borne out: the Geiger counters increased into an indistinguishable roar, and other instruments recording the neutron level on paper climbed continuously, doubling every two minutes until the reaction was shut down twenty-eight minutes later.

For this pile, $k$ was $1.0006$. On average, 0.6% of the neutrons emitted by a fissioning uranium atom are "delayed"—they are emitted by the further breakdown of short-lived fission products, rather than by the initial fission (the "prompt neutrons"). Thus the above pile had $k = 0.9946$ when considering only prompt neutrons, and its emissions increased on a slow exponential curve due to the contribution of delayed neutrons. A pile with $k = 1.0006$ for prompt neutrons would have doubled in neutron intensity every *tenth* of a second. If Fermi had not understood the atoms making up his pile and had only relied on its overall neutron-intensity graph to go on behaving like it had previously—or if he had just piled on uranium bricks, curious to observe empirically what would happen—then it would not have been a good year to be a student at the University of Chicago.

Nuclear weapons use conventional explosives to compress nuclear materials into a configuration with prompt $k \gg 1$; in a nuclear explosion, $k$ might be on the order of $2.3$, which is "vastly greater than one" for purposes of nuclear engineering.

At the time when the very first human-made critical reaction was initiated, Fermi already understood neutrons and uranium atoms—understood them sufficiently well to pull out the cadmium rod in careful increments, monitor the increasing reaction carefully, and shut it down after twenty-eight minutes. We do not currently have a strong grasp of the state space of cognitive algorithms. We do not have a strong grasp of how difficult or how easy it should be to improve cognitive problem-solving ability in a general AI by adding resources or trying to improve the underlying algorithms. We probably shouldn't expect to be able to do precise calculations; our state of uncertain knowledge about the space of cognitive algorithms probably shouldn't yield Fermi-style verdicts about when the trace will begin to climb without leveling off, down to a particular cadmium rod being pulled out twelve inches.

But we can hold out some hope of addressing larger, less exact questions, such as whether an AI trying to self-improve, or a global population of AIs trying to self-improve, can go "critical" ($k \approx 1^+$) or "supercritical" (prompt $k \gg 1$). We shouldn't

expect to predict exactly how many neutrons the metaphorical pile will output after two minutes. But perhaps we can predict in advance that piling on more and more uranium bricks will *eventually* cause the pile to start doubling its neutron production at a rate that grows quickly compared to its previous ascent . . . or, alternatively, conclude that self-modifying AIs should *not* be expected to improve at explosive rates.

So as not to allow this question to become too abstract, let us immediately consider some widely different stances that have been taken on the intelligence explosion debate. This is not an exhaustive list. As with any concrete illustration or "detailed storytelling," each case will import large numbers of auxiliary assumptions. I would also caution against labeling any particular case as "good" or "bad"—regardless of the true values of the unseen variables, we should try to make the best of them.

With those disclaimers stated, consider these concrete scenarios for a metaphorical "$k$ much less than one," "$k$ slightly more than one," and "prompt $k$ significantly greater than one," with respect to returns on cognitive investment.

**$k < 1$, the "intelligence fizzle":**

**Argument:** For most interesting tasks known to computer science, it requires exponentially greater investment of computing power to gain a linear return in performance. Most search spaces are exponentially vast, and low-hanging fruits are exhausted quickly. Therefore, an AI trying to invest an amount of cognitive work $w$ to improve its own performance will get returns that go as $\log(w)$, or if further reinvested, $\log(w + \log(w))$, and the sequence $\log(w)$, $\log(w + \log(w))$, $\log(w + \log(w + \log(w)))$ will converge very quickly.

**Scenario:** We might suppose that silicon intelligence is not significantly different from carbon, and that AI at the level of John von Neumann can be constructed, since von Neumann himself was physically realizable. But the constructed von Neumann does much less interesting work than the historical von Neumann, because the low-hanging fruits of science have already been exhausted. Millions of von Neumanns only accomplish logarithmically more work than one von Neumann, and it is not worth the cost of constructing such AIs. AI does not economically substitute for most cognitively skilled human labor, since even when smarter AIs can be built, humans can be produced more cheaply. Attempts are made to improve human intelligence via genetic engineering, or neuropharmaceuticals, or brain-computer interfaces, or cloning Einstein, etc.; but these attempts are foiled by the discovery that most "intelligence" is either unreproducible or not worth the cost of reproducing it. Moore's Law breaks down decisively, not just because of increasing technological difficulties of miniaturization, but because ever-faster computer chips don't accomplish much more than the previous generation of chips,

and so there is insufficient economic incentive for Intel to build new factories. Life continues mostly as before, for however many more centuries.

**$k \approx 1^+$, the "intelligence combustion":**

**Argument:** Over the last many decades, world economic growth has been roughly exponential—growth has neither collapsed below exponential nor exploded above, implying a metaphorical $k$ roughly equal to one (and slightly on the positive side). This is the characteristic behavior of a world full of smart cognitive agents making new scientific discoveries, inventing new technologies, and reinvesting resources to obtain further resources. There is no reason to suppose that changing from carbon to silicon will yield anything different. Furthermore, any single AI agent is unlikely to be significant compared to an economy of seven-plus billion humans. Thus AI progress will be dominated for some time by the contributions of the world economy to AI research, rather than by any one AI's internal self-improvement. No one agent is capable of contributing more than a tiny fraction of the total progress in computer science, and this doesn't change when human-equivalent AIs are invented.[5]

**Scenario:** The effect of introducing AIs to the global economy is a gradual, continuous increase in the overall rate of economic growth, since the first and most expensive AIs carry out a small part of the global economy's cognitive labor. Over time, the cognitive labor of AIs becomes cheaper and constitutes a larger portion of the total economy. The timescale of exponential growth starts out at the level of a human-only economy and gradually, continuously shifts to a higher growth rate—for example, Hanson (1998b) predicts world economic doubling times of between a month and a year. Economic dislocations are unprecedented but take place on a timescale which gives humans some chance to react.

**Prompt $k \gg 1$, the "intelligence explosion":**

**Argument:** The history of hominid evolution to date shows that it has not required exponentially greater amounts of evolutionary optimization to produce substantial real-world gains in cognitive performance—it did not require ten times the evolutionary interval to go from *Homo erectus* to *Homo sapiens* as from *Australopithecus* to *Homo erectus*.[6] All compound interest returned on discoveries such as the invention

5. I would attribute this rough view to Robin Hanson, although he hasn't confirmed that this is a fair representation.

6. This is incredibly oversimplified. See section 3.6 for a slightly less oversimplified analysis which ends up at roughly the same conclusion.

of agriculture, or the invention of science, or the invention of computers, has occurred without any ability of humans to reinvest technological dividends to increase their brain sizes, speed up their neurons, or improve the low-level algorithms used by their neural circuitry. Since an AI can reinvest the fruits of its intelligence in larger brains, faster processing speeds, and improved low-level algorithms, we should expect an AI's growth curves to be sharply above human growth curves.

**Scenario:** The first machine intelligence system to achieve sustainable returns on cognitive reinvestment is able to vastly improve its intelligence relatively quickly—for example, by rewriting its own software or by buying (or stealing) access to orders of magnitude more hardware on clustered servers. Such an AI is "prompt critical"—it can reinvest the fruits of its cognitive investments on short timescales, without the need to build new chip factories first. By the time such immediately accessible improvements run out, the AI is smart enough to, for example, crack the problem of protein structure prediction. The AI emails DNA sequences to online peptide synthesis labs (some of which boast a seventy-two-hour turnaround time), and uses the resulting custom proteins to construct more advanced ribosome equivalents (molecular factories). Shortly afterward, the AI has its own molecular nanotechnology and can begin construction of much faster processors and other rapidly deployed, technologically advanced infrastructure. This rough sort of scenario is sometimes colloquially termed "hard takeoff" or "AI-go-FOOM."[7]

There are many questions we could proceed to ask about these stances, which are actually points along a spectrum that compresses several different dimensions of potentially independent variance, etc. The implications from the arguments to the scenarios are also disputable. Further sections will address some of this in greater detail.

The broader idea is that different positions on "How large are the returns on cognitive reinvestment?" have widely different consequences with significant policy implications.

The problem of investing resources to gain more resources is fundamental in economics. An (approximately) rational agency will consider multiple avenues for improvement, purchase resources where they are cheapest, invest where the highest returns are expected, and try to bypass any difficulties that its preferences do not explicitly forbid

7. I must quickly remark that in my view, whether an AI attaining great power is a good thing or a bad thing would depend strictly on the AI's goal system. This in turn may depend on whether the programmers were able to solve the problem of "Friendly AI" (see Yudkowsky [2008a]).

This above point leads into another, different, and large discussion which is far beyond the scope of this paper, though I have very, *very* briefly summarized some core ideas in section 1.3. Nonetheless it seems important to raise the point that a hard takeoff/AI-go-FOOM scenario is not necessarily a bad thing, nor inevitably a good one.

bypassing. This is one factor that makes an artificial intelligence unlike a heap of uranium bricks: if you insert a cadmium-foil rod into a heap of uranium bricks, the bricks will not try to shove the rod back out, nor reconfigure themselves so that the rod absorbs fewer valuable neutrons. In economics, it is routine to suggest that a rational agency will do its best to overcome, bypass, or intelligently reconfigure its activities around an obstacle. Depending on the AI's preferences and capabilities, and on the surrounding society, it may make sense to steal poorly defended computing resources; returns on illegal investments are often analyzed in modern economic theory.

Hence the problem of describing an AI's curve for reinvested growth seems more like existing economics than existing problems in physics or computer science. As "microeconomics" is the discipline that considers rational agencies (such as individuals, firms, machine intelligences, and well-coordinated populations of machine intelligences) trying to maximize their returns on investment,[8] the posed open problem about growth curves under cognitive investment and reinvestment is titled "Intelligence Explosion Microeconomics."

Section 2 of this paper discusses the basic language for talking about the intelligence explosion and argues that we should pursue this project by looking for underlying microfoundations, not by pursuing analogies to allegedly similar historical events.

Section 3 attempts to showcase some specific informal reasoning about returns on cognitive investments, displaying the sort of arguments that have arisen in the context of the author explaining his stance on the intelligence explosion.

Section 4 proposes a tentative methodology for formalizing theories of the intelligence explosion—a project of describing possible microfoundations and explicitly stating their alleged relation to historical experience, such that some possibilities can be falsified.

Section 5 explores which subquestions seem both high value and possibly answerable. There are many things we'd like to know that we probably can't know given a reasonable state of uncertainty about the domain—for example, when will an intelligence explosion occur?

Section 6 summarizes and poses the open problem, and discusses what would be required for MIRI to fund further work in this area.

### 1.1. On (Extensionally) Defining Terms

It is obvious to ask questions like "What do you mean by 'intelligence'?" or "What sort of AI system counts as 'cognitively reinvesting'?" I shall attempt to answer these questions, but any definitions I have to offer should be taken as part of my own personal theory of the intelligence explosion. Consider the metaphorical position of early scientists

8. Academically, "macroeconomics" is about inflation, unemployment, monetary policy, and so on.

who have just posed the question "Why is fire hot?" Someone then proceeds to ask, "What exactly do you mean by 'fire'?" Answering, "Fire is the release of phlogiston" is presumptuous, and it is wiser to reply, "Well, for purposes of asking the question, fire is that bright orangey-red hot stuff coming out of that heap of sticks—which I think is really the release of phlogiston—but that definition is part of my answer, not part of the question itself."

I think it wise to keep this form of pragmatism firmly in mind when we are trying to define "intelligence" for purposes of analyzing the intelligence explosion.[9]

So as not to evade the question entirely, I usually use a notion of "intelligence ≡ efficient cross-domain optimization," constructed as follows:

1. Consider *optimization power* as the ability to steer the future into regions of possibility ranked high in a preference ordering. For instance, Deep Blue has the power to steer a chessboard's future into a subspace of possibility which it labels as "winning," despite attempts by Garry Kasparov to steer the future elsewhere. Natural selection can produce organisms much more able to replicate themselves than the "typical" organism that would be constructed by a randomized DNA string—evolution produces DNA strings that rank unusually high in fitness within the space of all DNA strings.[10]

2. Human cognition is distinct from bee cognition or beaver cognition in that human cognition is significantly more generally applicable across domains: bees build hives and beavers build dams, but a human engineer looks over both and then designs a dam with a honeycomb structure. This is also what separates Deep Blue, which only played chess, from humans, who can operate across many different domains and learn new fields.

9. On one occasion I was debating Jaron Lanier, who was arguing at length that it was bad to call computers "intelligent" because this would encourage human beings to act more mechanically, and therefore AI was impossible; and I finally said, "Do you mean to say that if I write a program and it writes a program and that writes another program and that program builds its own molecular nanotechnology and flies off to Alpha Centauri and starts constructing a Dyson sphere, that program is not *intelligent*?"

10. "Optimization" can be characterized as a concept we invoke when we expect a process to take on unpredictable intermediate states that will turn out to be apt for approaching a predictable destination—e.g., if you have a friend driving you to the airport in a foreign city, you can predict that your final destination will be the airport even if you can't predict any of the particular turns along the way. Similarly, Deep Blue's programmers retained their ability to predict Deep Blue's final victory by inspection of its code, even though they could not predict any of Deep Blue's particular moves along the way—if they knew exactly where Deep Blue would move on a chessboard, they would necessarily be at least that good at chess themselves.

3. Human engineering is distinct from natural selection, which is also a powerful cross-domain consequentialist optimizer, in that human engineering is faster and more computationally efficient. (For example, because humans can abstract over the search space, but that is a hypothesis about human intelligence, not part of my definition.)

In combination, these yield a definition of "intelligence ≡ efficient cross-domain optimization."

This tries to characterize "improved cognition" as the ability to produce solutions higher in a preference ordering, including, for example, a chess game with a higher probability of winning than a randomized chess game, an argument with a higher probability of persuading a human target, a transistor connection diagram that does more floating-point operations per second than a previous CPU, or a DNA string corresponding to a protein unusually apt for building a molecular factory. Optimization is characterized by an ability to hit narrow targets in a search space, where demanding a higher ranking in a preference ordering automatically narrows the measure of equally or more preferred outcomes. Improved intelligence is then hitting a narrower target in a search space, more computationally efficiently, via strategies that operate across a wider range of domains.

That definition is one which I invented for other purposes (my work on machine intelligence as such) and might not be apt for reasoning about the intelligence explosion. For purposes of discussing the intelligence explosion, it may be wiser to reason about forms of growth that more directly relate to quantities we can observe. The narrowness of the good-possibility space attained by a search process does not correspond very directly to most historical observables.

And for purposes of *posing the question* of the intelligence explosion, we may be better off with "Intelligence is that sort of *smartish stuff* coming out of brains, which can play chess, and price bonds, and persuade people to buy bonds, and invent guns, and figure out gravity by looking at wandering lights in the sky; and which, if a machine intelligence had it in large quantities, might let it invent molecular nanotechnology; and so on." To frame it another way, if something is powerful enough to build a Dyson Sphere, it doesn't really matter very much whether we call it "intelligent" or not. And this is just the sort of "intelligence" we're interested in—something powerful enough that whether or not we define it as "intelligent" is moot. This isn't to say that definitions are forbidden—just that further definitions would stake the further claim that those particular definitions were apt for carving reality at its joints, with respect to accurately predicting an intelligence explosion.

Choice of definitions has no power to affect physical reality. If you manage to define "AI self-improvement" in such a way as to exclude some smartish computer-thingy which carries out some mysterious internal activities on its own code for a week and

then emerges with a solution to protein structure prediction which it uses to build its own molecular nanotechnology . . . then you've obviously picked the wrong definition of "self-improvement." See, for example, the definition advocated by Mahoney (2010) in which "self-improvement" requires an increase in Kolmogorov complexity of an isolated system, or Bringsjord's (2012) definition in which a Turing machine is only said to self-improve if it can raise itself into a class of hypercomputers. These are both definitions which strike me as inapt for reasoning about the intelligence explosion, since it is not obvious (in fact I think it obviously false) that this sort of "self-improvement" is required to invent powerful technologies. One can define self-improvement to be the increase in Kolmogorov complexity of an isolated deterministic system, and proceed to prove that this can only go as the logarithm of time. But all the burden of showing that a real-world intelligence explosion is therefore impossible rests on the argument that doing impactful things in the real world requires an isolated machine intelligence to increase its Kolmogorov complexity. We should not fail to note that this is blatantly false.[11]

This doesn't mean that we should never propose more sophisticated definitions of self-improvement. It means we shouldn't lose sight of the wordless pragmatic background concept of an AI or AI population that rewrites its own code, or writes a successor version of itself, or writes an entirely new AI, or builds a better chip factory, or earns money to purchase more server time, or otherwise does something that increases the amount of pragmatically considered cognitive problem-solving capability sloshing around the system. And beyond that, "self-improvement" could describe genetically engineered humans, or humans with brain-computer interfaces, or upload clades, or several other possible scenarios of cognitive reinvestment, albeit here I will focus on the case of machine intelligence.[12]

It is in this spirit that I pose the open problem of formalizing I. J. Good's notion of the intelligence explosion. Coming up with good definitions for informal terms like "cognitive reinvestment," as they appear in the posed question, can be considered as part of the problem. In further discussion I suggest various definitions, categories, and

11. Since any system with a Kolmogorov complexity $k$ is unable to predict the Busy Beaver sequence for machines larger than $k$, increasing intelligence in the sense of being able to predict more of the Busy Beaver sequence would require increased Kolmogorov complexity. But since even galactic civilizations at Kardashev Level III probably can't predict the Busy Beaver sequence very far, limits on this form of "intelligence" are not very limiting. For more on this, see my informal remarks here.

12. This is traditional, but also sensible, since entirely computer-based, deliberately designed intelligences seem likely to be more apt for further deliberate improvement than biological brains. Biological brains are composed of giant masses of undocumented spaghetti code running on tiny noisy filaments that require great feats of medical ingenuity to read, let alone edit. This point is widely appreciated, but of course it is not beyond dispute.

distinctions. But such suggestions are legitimately disputable by anyone who thinks that a different set of definitions would be better suited to carving reality at its joints—to predicting what we will, in reality, actually observe to happen once some sort of smartish agency tries to invest in becoming smarterish.

### 1.2. Issues to Factor Out

Although we are ultimately interested only in the real-world results, I suggest that it will be productive theoretically—carve the issues at their natural joints—if we factor out for separate consideration issues of whether, for example, there might be an effective monitoring regime which could prevent an intelligence explosion, or whether the entire world economy will collapse due to global warming before then, and numerous other issues that don't seem to interact very strongly with the returns on cognitive investment *qua* cognitive investment.[13]

In particular, I would suggest explicitly factoring out all considerations of "What if an agent's preferences are such that it does not *want* to increase capability at the fastest rate it can achieve?" As Omohundro (2008) and Bostrom (2012) point out, most possible preferences imply capability increase as an instrumental motive. If you want to build an intergalactic civilization full of sentient beings leading well-lived lives, you will want access to energy and matter. The same also holds true if you want to fill space with two-hundred-meter giant cheesecakes. In either case you will also have an instrumental goal of becoming smarter. Just as you can fulfill most goals better by having access to more material resources, you can also accomplish more by being better at cognitive problems—by being able to hit narrower targets in a search space.

The space of all possible mind designs is vast (Muehlhauser and Salamon 2012), and there will always be *some* special case of an agent that chooses not to carry out any given deed (Armstrong, forthcoming). Given sufficient design competence, it should thus be possible to design an agent that doesn't prefer to ascend at the maximum possible rate—though expressing this within the AI's own preferences I would expect to be structurally nontrivial.

Even so, we need to separately consider the question of how fast a rational agency could intelligence-explode if it were trying to self-improve as fast as possible. If the maximum rate of ascent is already inherently slow, then there is little point in constructing a special AI design that prefers not to improve faster than its programmers can verify.

13. In particular, I would like to avoid round-robin arguments of the form "It doesn't matter if an intelligence explosion is possible, because there will be a monitoring regime that prevents it," and "It doesn't matter if the monitoring regime fails, because an intelligence explosion is impossible," where you never get to fully discuss either issue before being referred to the other side of the round-robin.

Policies are motivated by differentials of expected utility; there's no incentive to do any sort of action X intended to prevent Y unless we predict that Y might otherwise tend to follow assuming not-X. This requires us to set aside the proposed slowing factor and talk about what a rational agency might do if not slowed.

Thus I suggest that initial investigations of the intelligence explosion should consider the achievable rate of return on cognitive reinvestment for a rational agency trying to self-improve as fast as possible, in the absence of any obstacles not already present in today's world.[14] This also reflects the hope that trying to tackle the posed Open Problem should not require expertise in Friendly AI or international politics in order to talk about the returns on cognitive investment *qua* investment, even if predicting actual real-world outcomes might (or might not) require some of these issues to be factored back in.

### 1.3. AI Preferences: A Brief Summary of Core Theses

Despite the above, it seems impossible not to at least briefly summarize some of the state of discussion on AI preferences—if someone believes that a sufficiently powerful AI, or one which is growing at a sufficiently higher rate than the rest of humanity and hence gaining unsurpassable advantages, is unavoidably bound to kill everyone, then they may have a hard time dispassionately considering and analyzing the potential growth curves.

I have suggested that, in principle and in *difficult* practice, it should be possible to design a "Friendly AI" with programmer choice of the AI's preferences, and have the AI self-improve with sufficiently high fidelity to knowably keep these preferences stable. I also think it should be possible, in principle and in difficult practice, to convey the complicated information inherent in human preferences into an AI, and then apply further idealizations such as reflective equilibrium and ideal advisor theories (Muehlhauser and Williamson 2013) so as to arrive at an output which corresponds intuitively to the AI "doing the right thing." See also Yudkowsky (2008a).

On a larger scale the current state of discussion around these issues seems to revolve around four major theses:

The *Intelligence Explosion Thesis* says that, due to recursive self-improvement, an AI can potentially grow in capability on a timescale that seems fast relative to human experience. This in turn implies that strategies which rely on humans reacting to and

14. That is, we might assume that people continue to protect their home computers with firewalls, for whatever that is worth. We should not assume that there is a giant and effective global monitoring organization devoted to stamping out any sign of self-improvement in AIs à la the Turing Police in William Gibson's (1984) *Neuromancer*. See also the sort of assumptions used in Robert Freitas's (2000) *Some Limits to Global Ecophagy*, wherein proposed limits on how fast the biosphere can be converted into nanomachines revolve around the assumption that there is a global monitoring agency looking for unexplained heat blooms, and that this will limit the allowed heat dissipation of nanomachines.

restraining or punishing AIs are unlikely to be successful in the long run, and that what the first strongly self-improving AI prefers can end up mostly determining the final outcomes for Earth-originating intelligent life. (This subthesis is the entire topic of the current paper. One observes that the arguments surrounding the thesis are much more complex than the simple summary above would suggest. This is also true of the other three theses below.)

The *Orthogonality Thesis* says that mind-design space is vast enough to contain minds with almost any sort of preferences. There exist instrumentally rational agents which pursue almost any utility function, and they are mostly stable under reflection. See Armstrong (forthcoming) and Muehlhauser and Salamon (2012). There are many strong arguments for the Orthogonality Thesis, but one of the strongest proceeds by construction: If it is possible to answer the purely epistemic question of which actions would lead to how many paperclips existing, then a paperclip-seeking agent is constructed by hooking up that answer to motor output. If it is very good at answering the epistemic question of which actions would result in great numbers of paperclips, then it will be a very instrumentally powerful agent.[15]

The *Complexity of Value Thesis* says that human values are complex in the sense of having high algorithmic (Kolmogorov) complexity (Yudkowsky 2011; Muehlhauser and Helm 2012). Even idealized forms of human value, such as reflective equilibrium (Rawls 1971) or ideal advisor theories (Rosati 1995)—what we *would* want in the limit of infinite knowledge of the world, infinite thinking speeds, and perfect self-understanding, etc.—are predicted to still have high algorithmic complexity. This tends to follow from naturalistic theories of metaethics under which human preferences for happiness, freedom, growth, aesthetics, justice, etc., (see Frankena [1973, chap. 5] for one list of commonly stated terminal values) have no privileged reason to be readily reducible to each other or to anything else. The Complexity of Value Thesis is that to realize valuable outcomes, an AI must have complex information in its utility function; it also will

15. Such an agent will not modify itself to seek something else, because this would lead to fewer paperclips existing in the world, and its criteria for all actions including internal actions is the number of expected paperclips. It will not modify its utility function to have properties that humans would find more pleasing, because it does not already care about such metaproperties and is not committed to the belief that paperclips occupy a maximum of such properties; it is an expected *paperclip* maximizer, not an expected *utility* maximizer.

Symmetrically, AIs which have been successfully constructed to start with "nice" preferences in their initial state will not throw away those nice preferences merely in order to confer any particular logical property on their utility function, unless they were already constructed to care about that property.

not suffice to tell it to "just make humans happy" or any other simplified, compressed principle.[16]

The *Instrumental Convergence Thesis* says that for most choices of a utility function, instrumentally rational agencies will predictably wish to obtain certain generic resources, such as matter and energy, and pursue certain generic strategies, such as not making code changes which alter their effective future preferences (Omohundro 2008; Bostrom 2012). Instrumental Convergence implies that an AI does not need to have specific terminal values calling for it to harm humans, in order for humans to be harmed. The AI does not hate you, but neither does it love you, and you are made of atoms that it can use for something else.

In combination, the Intelligence Explosion Thesis, the Orthogonality Thesis, the Complexity of Value Thesis, and the Instrumental Convergence Thesis imply a very large utility differential for whether or not we can solve the design problems (1) relating to a self-improving AI with stable specifiable preferences and (2) relating to the successful transfer of human values (and their further idealization via, e.g., reflective equilibrium or ideal advisor theories), with respect to the *first* AI to undergo the intelligence explosion.

All this is another and quite different topic within the larger discussion of the intelligence explosion, compared to its microeconomics. Here I will only note that large returns on cognitive investment need not correspond to unavoidable horror scenarios so painful that we are forced to argue against them, nor to virtuous pro-science-and-technology scenarios that virtuous people ought to affiliate with. For myself I would tend to view larger returns on cognitive reinvestment as corresponding to increased policy-dependent variance. And whatever the true values of the unseen variables, the question is not whether they sound like "good news" or "bad news"; the question is how we can improve outcomes as much as possible given those background settings.

## 2. Microfoundations of Growth

Consider the stance on the intelligence explosion thesis which says: "I think we should expect that exponentially greater investments—of computing hardware, software programming effort, etc.—will only produce linear gains in real-world performance on cognitive tasks, since most search spaces are exponentially large. So the fruits of machine

16. The further arguments supporting the Complexity of Value suggest that even "cosmopolitan" or "non-human-selfish" outcomes have implicit specifications attached of high Kolmogorov complexity. Perhaps you would hold yourself to be satisfied with a future intergalactic civilization full of sentient beings happily interacting in ways you would find incomprehensible, even if none of them are you or human-derived. But an expected paperclip maximizer would fill the galaxies with paperclips instead. This is why expected paperclip maximizers are scary.

intelligence reinvested into AI will only get logarithmic returns on each step, and the 'intelligence explosion' will peter out very quickly."

Is this scenario plausible or implausible? Have we seen anything in the real world—made any observation, ever—that should affect our estimate of its probability?

*(At this point, I would suggest that the serious reader turn away and take a moment to consider this question on their own before proceeding.)*

Some possibly relevant facts might be:

- Investing exponentially more computing power into a constant chess-playing program produces linear increases in the depth of the chess-game tree that can be searched, which in turn seems to correspond to linear increases in Elo rating (where two opponents of a fixed relative Elo distance, regardless of absolute ratings, theoretically have a constant probability of losing or winning to each other).

- Chess-playing algorithms have recently improved much faster than chess-playing hardware, particularly since chess-playing programs began to be open-sourced. Deep Blue ran on 11.8 billion floating-point operations per second and had an Elo rating of around 2,700; Deep Rybka 3 on a Intel Core 2 Quad 6600 has an Elo rating of 3,202 on 2.4 billion floating-point operations per second.[17]

- It seems that in many important senses, humans get more than four times the real-world return on our intelligence compared to our chimpanzee cousins. This was achieved with *Homo sapiens* having roughly four times as much cortical volume and six times as much prefrontal cortex.[18]

- Within the current human species, measured IQ is entangled with brain size; and this entanglement is around a 0.3 correlation in the variances, rather than, say, a doubling of brain size being required for each ten-point IQ increase.[19]

- The various Moore's-like laws measuring computing technologies, operations per second, operations per dollar, disk space per dollar, and so on, are often said to have characteristic doubling times ranging from twelve months to three years; they are formulated so as to be exponential with respect to time. People have written papers

17. Score determined (plus or minus ∼23) by the Swedish Chess Computer Association based on 1,251 games played on the tournament level.

18. The obvious conclusion you might try to draw about hardware scaling is oversimplified and would be relevantly wrong. See section 3.1.

19. For entrants unfamiliar with modern psychological literature: Yes, there is a strong correlation $g$ between almost all measures of cognitive ability, and IQ tests in turn are strongly correlated with this $g$ factor and well correlated with many measurable life outcomes and performance measures. See the Cambridge Handbook of Intelligence (Sternberg and Kaufman 2011).

questioning Moore's Law's validity (see, e.g., Tuomi [2002]); and the Moore's-like law for serial processor speeds broke down in 2004. The original law first observed by Gordon Moore, over transistors per square centimeter, has remained on track.

- Intel has invested exponentially more researcher-hours and inflation-adjusted money to invent the technology and build the manufacturing plants for successive generations of CPUs. But the CPUs themselves are increasing exponentially in transistor operations per second, not linearly; and the computer-power doubling time is shorter (that is, the exponent is higher) than that of the increasing investment cost.[20]

- The amount of evolutionary time (a proxy measure of cumulative selection pressure and evolutionary optimization) which produced noteworthy changes during human and hominid evolution does not seem to reveal exponentially greater amounts of time invested. It did not require ten times as long to go from *Homo erectus* to *Homo sapiens*, as from *Australopithecus* to *Homo erectus*.[21]

- World economic output is roughly exponential and increases faster than population growth, which is roughly consistent with exponentially increasing investments producing exponentially increasing returns. That is, roughly linear (but with multiplication factor $k > 1$) returns on investment. On a larger timescale, world-historical economic output can be characterized as a sequence of exponential modes (Hanson 1998b). Total human economic output was also growing exponentially in AD 1600 or 2000 BC, but with smaller exponents and much longer doubling times.

- Scientific output in "total papers written" tends to grow exponentially with a short doubling time, both globally (around twenty-seven years [NSB 2012, chap. 5]) and within any given field. But it seems extremely questionable whether there has been more global change from 1970 to 2010 than from 1930 to 1970. (For readers who have heard relatively more about "accelerating change" than about "the Great Stagnation": the claim is that total-factor productivity growth in, e.g., the United States dropped from 0.75% per annum before the 1970s to 0.25% thereafter [Cowen 2011].) A true cynic might claim that, in many fields, exponentially greater

20. As Carl Shulman observes, Intel does not employ 343 million people.

21. One might ask in reply whether *Homo erectus* is being singled out on the basis of being distant enough in time to have its own species name, rather than by any prior measure of cognitive ability. This issue is taken up at much greater length in section 3.6.

investment in science is yielding a roughly constant amount of annual progress—sublogarithmic returns![22]

- This graph (Silver 2012) shows how many books were authored in Europe as a function of time; after the invention of the printing press, the graph jumps in a sharp, faster-than-exponential upward surge.
- All technological progress in known history has been carried out by essentially constant human brain architectures. There are theses about continuing human evolution over the past ten thousand years, but all such changes are nowhere near the scale of altering "You have a brain that's more or less 1,250 cubic centimeters of dendrites and axons, wired into a prefrontal cortex, a visual cortex, a thalamus, and so on." It has not required much larger brains, or much greater total cumulative selection pressures, to support the continuing production of more sophisticated technologies and sciences over the human regime.
- The amount of complex order per unit time created by a human engineer is completely off the scale compared to the amount of complex order per unit time created by natural selection within a species. A single mutation conveying a 3% fitness advantage would be expected to take 768 generations to rise to fixation through a sexually reproducing population of a hundred thousand members. A computer programmer can design new complex mechanisms with hundreds of interoperating parts over the course of a day or an hour. In turn, the amount of complex order per unit time created by natural selection is completely off the scale for Earth before the dawn of life. A graph of "order created per unit time" during Earth's history would contain two discontinuities representing the dawn of fundamentally different optimization processes.

The list of observations above might give you the impression that it could go either way—that some things are exponential and some things aren't. Worse, it might look like an invitation to decide your preferred beliefs about AI self-improvement as a matter of emotional appeal or fleeting intuition, and then decide that any of the above cases which behave similarly to how you think AI self-improvement should behave, are the natural historical examples we should consult to determine the outcome of AI. For example, clearly the advent of self-improving AI seems most similar to other economic

22. I am in fact such a true cynic and I suspect that social factors dilute average contributions around as fast as new researchers can be added. A less cynical hypothesis would be that earlier science is easier, and later science grows more difficult at roughly the same rate that scientific output scales with more researchers being added.

speedups like the invention of agriculture.[23] Or obviously it's analogous to other foundational changes in the production of complex order, such as human intelligence or self-replicating life.[24] Or self-evidently the whole foofaraw is analogous to the panic over the end of the Mayan calendar in 2012 since it belongs in the reference class of "supposed big future events that haven't been observed."[25] For more on the problem of "reference class tennis," see section 2.1.

It seems to me that the real lesson to be derived from the length of the above list is that we shouldn't expect some single grand law about whether you get superexponential, exponential, linear, logarithmic, or constant returns on cognitive investments. The cases above have different behaviors; they are not all conforming to a single Grand Growth Rule.

It's likewise not the case that Reality proceeded by randomly drawing a curve type from a barrel to assign to each of these scenarios, and the curve type of "AI self-improvement" will be independently sampled with replacement from the same barrel. So it likewise doesn't seem valid to argue about how likely it is that someone's personal favorite curve type gets drawn by trumpeting historical cases of that curve type, thereby proving that it's more frequent within the Curve Type Barrel and more likely to be randomly drawn.

Most of the processes cited above yielded fairly regular behavior over time. Meaning that the attached curve was actually characteristic of that process's causal mechanics, and a predictable feature of those mechanics, rather than being assigned and reassigned at random. Anyone who throws up their hands and says, "It's all unknowable!" may also be scoring fewer predictive points than they could.

These differently behaving cases are not competing arguments about how a single grand curve of cognitive investment has previously operated. They are all simultaneously true, and hence they must be telling us *different* facts about growth curves—telling us about different domains of a multivariate growth function—advising us of many compatible truths about how intelligence and real-world power vary with different kinds of cognitive investments.[26]

---

23. See Hanson (2008a).

24. See Yudkowsky (2008b, 2008c, 2008d).

25. See, e.g., this post in an online discussion.

26. Reality itself is always perfectly consistent—only maps can be in conflict, not the territory. Under the Bayesian definition of evidence, "strong evidence" is just that sort of evidence that we almost never see on more than one side of an argument. Unless you've made a mistake somewhere, you should almost never see extreme likelihood ratios pointing in different directions. Thus it's not possible that the facts listed are all "strong" arguments, about the *same* variable, pointing in *different* directions.

Rather than selecting one particular historical curve to anoint as characteristic of the intelligence explosion, it might be possible to build an underlying causal model, one which would be compatible with all these separate facts. I would propose that we should be trying to formulate a microfoundational model which, rather than just generalizing over surface regularities, tries to describe underlying causal processes and returns on particular types of cognitive investment. For example, rather than just talking about how chess programs have improved over time, we might try to describe how chess programs improve as a function of computing resources plus the cumulative time that human engineers spend tweaking the algorithms. Then in turn we might say that human engineers have some particular *intelligence* or *optimization power*, which is different from the optimization power of a chimpanzee or the processes of natural selection. The process of building these causal models would hopefully let us arrive at a more realistic picture—one compatible with the many different growth curves observed in different historical situations.

### 2.1. The Outside View versus the Lucas Critique

A fundamental tension in the so-far-informal debates on intelligence explosion has been the rough degree of abstraction that is trustworthy and useful when modeling these future events.

The first time I happened to occupy the same physical room as Ray Kurzweil, I asked him why his graph of Moore's Law showed the events of "a $1,000 computer is as powerful as a human brain," "a $1,000 computer is a thousand times as powerful as a human brain," and "a $1,000 computer is a billion times as powerful as a human brain," all following the same historical trend of Moore's Law.[27] I asked, did it really make sense to continue extrapolating the humanly observed version of Moore's Law past the point where there were putatively minds with a billion times as much computing power?

$\text{Kurzweil}_{2001}$ replied that the existence of machine superintelligence was exactly what would provide the fuel for Moore's Law to continue and make it possible to keep developing the required technologies. In other words, $\text{Kurzweil}_{2001}$ regarded Moore's Law as the primary phenomenon and considered machine superintelligence a secondary

27. The same chart showed allegedly "human-level computing power" as the threshold of predicted AI, which is a methodology I strongly disagree with, but I didn't want to argue with that part at the time. I've looked around in Google Images for the exact chart but didn't find it; Wikipedia does cite similar predictions as having been made in *The Age of Spiritual Machines* (Kurzweil 1999), but Wikipedia's cited timelines are shorter term than I remember.

phenomenon which ought to assume whatever shape was required to keep the primary phenomenon on track.[28]

You could even imagine arguing (though $\text{Kurzweil}_{2001}$ did not say this part) that we've seen Moore's Law continue through many generations and across many different types of hardware, while we have no actual experience with machine superintelligence. So an extrapolation of Moore's Law should take epistemic primacy over more speculative predictions about superintelligence because it's based on more experience and firmer observations.

My own interpretation of the same history would be that there was some underlying difficulty curve for how more sophisticated CPUs required more knowledge and better manufacturing technology to build, and that over time human researchers exercised their intelligence to come up with inventions, tools to build more inventions, physical theories, experiments to test those theories, programs to help design CPUs,[29] etc. The process whereby more and more transistors are packed into a given area every eighteen months should not be an exogenous factor of how often the Earth traverses 1.5 orbits around the Sun; it should be a function of the engineers. So if we had *faster engineers*, we would expect a faster form of Moore's Law. (See section 3.3 for related points and counterpoints about fast manipulator technologies and sensor bandwidth also being required.)

$\text{Kurzweil}_{2001}$ gave an impromptu response seeming to suggest that Moore's Law might become more difficult at the same rate that superintelligence increased in problem-solving ability, thus preserving the forecast for Moore's Law in terms of time. But why should that be true? We don't have an exact idea of what the historical intrinsic-difficulty curve looked like; it's difficult to observe directly. Our main data is the much-better-known Moore's Law trajectory which describes how fast human engineers were able to traverse the difficulty curve over outside time.[30] But we could still reasonably expect that, if our old extrapolation was for Moore's Law to follow such-and-such curve given human engineers, then faster engineers should break upward from that extrapolation.

---

28. I attach a subscript by year because (1) Kurzweil was replying on the spot so it is not fair to treat his off-the-cuff response as a permanent feature of his personality and (2) Sandberg (2010) suggests that Kurzweil has changed his position since then.

29. There are over two billion transistors in the largest Core i7 processor. At this point human engineering *requires* computer assistance.

30. One can imagine that Intel may have balanced the growth rate of its research investments to follow industry expectations for Moore's Law, even as a much more irregular underlying difficulty curve became steeper or shallower. This hypothesis doesn't seem inherently untestable—someone at Intel would actually have had to make those sorts of decisions—but it's not obvious to me how to check it on previously gathered, easily accessed data.

Or to put it more plainly, the fully-as-naive extrapolation in the other direction would be, "Given human researchers of constant speed, computing speeds double every 18 months. So if the researchers are running on computers themselves, we should expect computing speeds to double in 18 months, then double again in 9 physical months (or 18 subjective months for the 2x-speed researchers), then double again in 4.5 physical months, and finally reach infinity after a total of 36 months." If humans accumulate subjective time at a constant rate $x = t$, and we observe that computer speeds increase as a Moore's-Law exponential function of subjective time $y = e^x$, then when subjective time increases at the rate of current computer speeds we get the differential equation $y' = e^y$ whose solution has computer speeds increasing hyperbolically, going to infinity after finite time.[31] (See, e.g., the model of Moravec [1999].)

In real life, we might not believe this as a quantitative estimate. We might not believe that in real life such a curve would have, even roughly, a hyperbolic shape before it started hitting (high) physical bounds. But at the same time, we might in real life believe that research ought to go substantially faster if the researchers could reinvest the fruits of their labor into their own cognitive speeds—that we are seeing an important hint buried within this argument, even if its details are wrong. We could believe as a qualitative prediction that "if computer chips are following Moore's Law right now with human researchers running at constant neural processing speeds, then in the hypothetical scenario where the researchers are running on computers, we should see a new Moore's Law bounded far below by the previous one." You might say something like, "Show me a reasonable model of how difficult it is to build chips as a function of knowledge, and how knowledge accumulates over subjective time, and you'll get a hyperexponential explosion out of Moore's Law once the researchers are running on computers. Conversely, if you give me a regular curve of increasing difficulty which *averts* an intelligence explosion, it will falsely retrodict that human engineers should only be able to get subexponential improvements out of computer technology. And of course it would be *unreasonable*—a specific unsupported miraculous irregularity of the curve—for making chips to suddenly get much more difficult to build, coincidentally exactly as AIs started doing research. The difficulty curve might shift upward at some random later point, but there'd still be a bonanza from whatever improvement was available up until then."

In turn, that reply gets us into a rather thorny meta-level issue:

> A: Why are you introducing all these strange new *unobservable* abstractions? We can see chips getting faster over time. That's what we can measure and that's what we have experience with. Who measures this *difficulty* of

31. The solution of $dy/dt = e^y$ is $y = -\log(c - t)$ and $dy/dt = 1/(c - t)$.

> which you speak? Who measures *knowledge*? These are all made-up quantities with no rigorous basis in reality. What we do have solid observations of is the number of transistors on a computer chip, per year. So I'm going to project that extremely regular curve out into the future and extrapolate from there. The rest of this is sheer, loose speculation. Who knows how many other possible supposed "underlying" curves, besides this "knowledge" and "difficulty" business, would give entirely different answers?

To which one might reply:

> B: Seriously? Let's consider an extreme case. Neurons spike around 2–200 times per second, and axons and dendrites transmit neural signals at 1–100 meters per second, less than a millionth of the speed of light. Even the heat dissipated by each neural operation is around six orders of magnitude above the thermodynamic minimum at room temperature.[32] Hence it should be physically possible to speed up "internal" thinking (which doesn't require "waiting on the external world") by at least six orders of magnitude without resorting to smaller, colder, reversible, or quantum computers. Suppose we were dealing with minds running a million times as fast as a human, at which rate they could do a year of internal thinking in thirty-one seconds, such that the total subjective time from the birth of Socrates to the death of Turing would pass in 20.9 hours. Do you still think the best estimate for how long it would take them to produce their next generation of computing hardware would be 1.5 orbits of the Earth around the Sun?

Two well-known epistemological stances, with which the respective proponents of these positions could identify their arguments, would be the *outside view* and the *Lucas critique*.

---

32. The brain as a whole organ dissipates around 20 joules per second, or 20 watts. The minimum energy required for a one-bit irreversible operation (as a function of temperature $T$) is $kT \ln(2)$, where $k = 1.38 \cdot 10^{23}$ joules/kelvin is Boltzmann's constant, and $\ln(2)$ is the natural log of 2 (around 0.7). Three hundred kelvin is 27°C or 80°F. Thus under ideal circumstances 20 watts of heat dissipation corresponds to $7 \cdot 10^{21}$ irreversible binary operations per second at room temperature.

The brain can be approximated as having $10^{14}$ synapses. I found data on average synaptic activations per second hard to come by, with different sources giving numbers from 10 activations per second to 0.003 activations/second (not all dendrites must activate to trigger a spike, and not all neurons are highly active at any given time). If we approximate the brain as having $10^{14}$ synapses activating on the order of once per second on average, this would allow $\sim 10^2$ irreversible operations per synaptic activation after a $10^6$-fold speedup.

(Note that since each traveling impulse of electrochemical activation requires many chemical ions to be pumped back across the neuronal membrane afterward to reset it, total distance traveled by neural impulses is a more natural measure of expended biological energy than total activations. No similar rule would hold for photons traveling through optical fibers.)

The "outside view" (Kahneman and Lovallo 1993) is a term from the heuristics and biases program in experimental psychology. A number of experiments show that if you ask subjects for estimates of, say, when they will complete their Christmas shopping, the right question to ask is, "When did you finish your Christmas shopping last year?" and not, "How long do you think it will take you to finish your Christmas shopping?" The latter estimates tend to be vastly over-optimistic, and the former rather more realistic. In fact, as subjects are asked to make their estimates using more detail—visualize where, when, and how they will do their Christmas shopping—their estimates become more optimistic, and less accurate. Similar results show that the actual planners and implementers of a project, who have full acquaintance with the internal details, are often much more optimistic and much less accurate in their estimates compared to experienced outsiders who have relevant experience of similar projects but don't know internal details. This is sometimes called the dichotomy of the *inside view* versus the *outside view*. The "inside view" is the estimate that takes into account all the details, and the "outside view" is the very rough estimate that would be made by comparing your project to other roughly similar projects without considering any special reasons why this project might be different.

The *Lucas critique* (Lucas 1976) in economics was written up in 1976 when "stagflation"—simultaneously high inflation and unemployment—was becoming a problem in the United States. Robert Lucas's concrete point was that the Phillips curve trading off unemployment and inflation had been observed at a time when the Federal Reserve was trying to moderate inflation. When the Federal Reserve gave up on moderating inflation in order to drive down unemployment to an even lower level, employers and employees adjusted their long-term expectations to take into account continuing inflation, and the Phillips curve shifted.

Lucas's larger and meta-level point was that the previously observed Phillips curve wasn't fundamental enough to be *structurally invariant* with respect to Federal Reserve policy—the concepts of inflation and unemployment weren't deep enough to describe elementary things that would remain stable even as Federal Reserve policy shifted. A very succinct summary appears in Wikipedia (2013):

> The Lucas critique suggests that if we want to predict the effect of a policy experiment, we should model the "deep parameters" (relating to preferences, technology and resource constraints) that are assumed to govern *individual* behavior; so called "microfoundations." If these models can account for observed empirical regularities, we can then predict what individuals will do, *taking into account* the change in policy, and then aggregate the individual decisions to calculate the macroeconomic effects of the policy change.

The main explicit proponent of the outside view in the intelligence explosion debate is Robin Hanson, who also proposes that an appropriate reference class into which to place the "Singularity"—a term not specific to the intelligence explosion but sometimes including it—would be the reference class of major economic transitions resulting in substantially higher exponents of exponential growth. From Hanson's (2008a) blog post "Outside View of Singularity":

> Most everything written about a possible future singularity takes an inside view, imagining details of how it might happen. Yet people are seriously biased toward inside views, forgetting how quickly errors accumulate when reasoning about details. So how far can we get with an outside view of the next singularity?
>
> Taking a long historical long view, we see steady total growth rates punctuated by rare transitions when new faster growth modes appeared with little warning. We know of perhaps four such "singularities": animal brains (∼600 MYA), humans (∼2 MYA), farming (∼10 KYA), and industry (∼0.2 KYA). The statistics of previous transitions suggest we are perhaps overdue for another one, and would be substantially overdue in a century. The next transition would change the growth rate rather than capabilities directly, would take a few years at most, and the new doubling time would be a week to a month.

More on this analysis can be found in Hanson (1998b).

The original blog post concludes:

> Excess inside viewing usually continues even after folks are warned that outside viewing works better; after all, inside viewing better shows off inside knowledge and abilities. People usually justify this via reasons why the current case is exceptional. (Remember how all the old rules didn't apply to the new dotcom economy?) So expect to hear excuses why the next singularity is also an exception where outside view estimates are misleading. Let's keep an open mind, but a wary open mind.

Another of Hanson's (2008c) posts, in what would later be known as the Yudkowsky-Hanson AI-Foom Debate, said:

> It is easy, way too easy, to generate new mechanisms, accounts, theories, and abstractions. To see if such things are *useful*, we need to vet them, and that is easiest "nearby," where we know a lot. When we want to deal with or understand things "far," where we know little, we have little choice other than to rely on mechanisms, theories, and concepts that have worked well near. Far is just the wrong place to try new things.

> There are a bazillion possible abstractions we could apply to the world. For each abstraction, the question is not whether one *can* divide up the world that way, but whether it "carves nature at its joints," giving *useful* insight not easily gained via other abstractions. We should be wary of inventing new abstractions just to make sense of things far; we should insist they first show their value nearby.

The lesson of the outside view pushes us to use abstractions and curves that are clearly empirically measurable, and to beware inventing new abstractions that we can't see directly.

The lesson of the Lucas critique pushes us to look for abstractions deep enough to describe growth curves that would be stable in the face of minds improving in speed, size, and software quality.

You can see how this plays out in the tension between "Let's predict computer speeds using this very well-measured curve for Moore's Law over time—where the heck is all this other stuff coming from?" versus "But almost any reasonable causal model that describes the role of human thinking and engineering in producing better computer chips, ought to predict that Moore's Law would speed up once computer-based AIs were carrying out all the research!"

It would be unfair to use my passing exchange with Kurzweil as a model of the debate between myself and Hanson. Still, I did feel that the basic disagreement came down to a similar tension—that Hanson kept raising a skeptical and unmoved eyebrow at the wild-eyed, empirically unvalidated, complicated abstractions which, from my perspective, constituted my attempt to put *any* sort of microfoundations under surface curves that couldn't possibly remain stable.

Hanson's overall prototype for visualizing the future was an economic society of *ems*, software emulations of scanned human brains. It would then be possible to turn capital inputs (computer hardware) into skilled labor (copied ems) almost immediately. This was Hanson's explanation for how the em economy could follow the "same trend" as past economic speedups, to a world economy that doubled every year or month (vs. a roughly fifteen-year doubling time at present [Hanson 1998b]).

I thought that the idea of copying human-equivalent minds missed almost every potentially interesting aspect of the intelligence explosion, such as faster brains, larger brains, or above all better-designed brains, all of which seemed liable to have far greater effects than increasing the quantity of workers.

Why? That is, if you can invest a given amount of computing power in more brains, faster brains, larger brains, or improving brain algorithms, why think that the return on investment would be significantly higher in one of the latter three cases?

A more detailed reply is given in section 3, but in quick summary:

There's a saying in software development, "Nine women can't have a baby in one month," meaning that you can't get the output of ten people working for ten years by hiring a hundred people to work for one year, or more generally, that working time scales better than the number of people, *ceteris paribus*. It's also a general truth of computer science that fast processors can simulate parallel processors but not always the other way around. Thus we'd expect the returns on speed to be higher than the returns on quantity.

We have little solid data on how human intelligence scales with added neurons and constant software. Brain size does vary between humans and this variance correlates by about 0.3 with $g$ (McDaniel 2005), but there are reams of probable confounders, such as childhood nutrition. Humans have around four times the brain volume of chimpanzees, but the difference between us is probably mostly brain-level cognitive algorithms.[33] It is a general truth of computer science that if you take one processing unit and split it up into ten parts with limited intercommunication bandwidth, they can do no better than the original on any problem, and will do considerably worse on many problems. Similarly we might expect that, for most intellectual problems, putting on ten times as many researchers running human software scaled down to one-fifth the brain size would probably not be a net gain, and that, for many intellectual problems, researchers with four times the brain size would probably be a significantly greater gain than adding four times as many researchers.[34]

Trying to say how intelligence and problem-solving ability scale with improved cognitive algorithms is even harder to relate to observation. In any computer-based field where surface capabilities are visibly improving, it is usually true that you are better off with modern algorithms and a computer from ten years earlier, compared to a modern computer and the algorithms from ten years earlier. This is definitely true in computer chess, even though the net efforts put in by chess-program enthusiasts to create better programs are small compared to the vast effort Intel puts into creating better computer chips every year. But this observation only conveys a small fraction of the idea that you can't match a human's intellectual output using any number of chimpanzees.

Informally, it looks to me like

$$\text{quantity} < (\text{size}, \text{speed}) < \text{quality}$$

when it comes to minds.

Hanson's scenario in which all investments went into increasing the mere quantity of ems—and this was a good estimate of the total impact of an intelligence explosion—

33. If it were possible to create a human just by scaling up an *Australopithecus* by a factor of four, the evolutionary path from *Australopithecus* to us would have been much shorter.

34. Said with considerable handwaving. But do you really think that's false?

seemed to imply that the returns on investment from larger brains, faster thinking, and improved brain designs could all be neglected, which implied that the returns from such investments were relatively low.[35] Whereas it seemed to me that any reasonable microfoundations which were compatible with prior observation—which didn't retrodict that a human should be intellectually replaceable by ten chimpanzees—should imply that quantity of labor wouldn't be the dominating factor. Nonfalsified growth curves ought to say that, given an amount of computing power which you could invest in more minds, faster minds, larger minds, or better-designed minds, you would invest in one of the latter three.

We don't invest in larger human brains because that's impossible with current technology—we can't just hire a researcher with three times the cranial volume, we can only throw more warm bodies at the problem. If that investment avenue suddenly became available . . . it would probably make quite a large difference, pragmatically speaking. I was happy to concede that my model only made vague qualitative predictions—I didn't think I had enough data to make quantitative predictions like Hanson's estimates of future economic doubling times. But qualitatively I thought it obvious that all these hard-to-estimate contributions from faster brains, larger brains, and improved underlying cognitive algorithms were all pointing along the same rough vector, namely "way up." Meaning that Hanson's estimates, sticking to extrapolated curves of well-observed quantities, would be predictably biased way down.

Whereas from Hanson's perspective, this was all wild-eyed unverified speculation, and he was sticking to analyzing ems because we had a great deal of data about how human minds worked and no way to solidly ground all these new abstractions I was hypothesizing.

Aside from the Lucas critique, the other major problem I have with the "outside view" is that everyone who uses it seems to come up with a different reference class and a different answer. To Ray Kurzweil, the obvious reference class for "the Singularity" is Moore's Law as it has operated over recent history, not Hanson's comparison to agriculture. In this post an online discussant of these topics places the "Singularity" into the reference class "beliefs in coming of a new world" which has "a 0% success rate" . . . explicitly terming this the proper "outside view" of the situation using "reference class forecasting," and castigating anyone who tried to give a different answer as having used an "inside view." For my response to all this at greater length, see "'Outside View!'

35. Robin Hanson replied to a draft of this paper: "The fact that I built a formal model that excluded these factors doesn't mean I think such effects are so small as to be negligible. Not only is it reasonable to build models that neglect important factors, it is usually impossible not to do so." This is surely true; nonetheless, I think that in this case the result was a predictable directional bias.

as Conversation-Halter" (Yudkowsky 2010). The gist of my reply was that the outside view has been experimentally demonstrated to beat the inside view for software projects that are similar to previous software projects, and for this year's Christmas shopping, which is highly similar to last year's Christmas shopping. The outside view would be expected to work less well on a new thing that is less similar to the old things than all the old things were similar to each other—especially when you try to extrapolate from one kind of causal system to a very different causal system. And one major sign of trying to extrapolate across too large a gap is when everyone comes up with a different "obvious" reference class.

Of course it also often happens that disputants think different microfoundations—different causal models of reality—are "obviously" appropriate. But then I have some idea of how to zoom in on hypothesized causes, assess their simplicity and regularity, and figure out how to check them against available evidence. I don't know what to do after two people take different reference classes and come up with different outside views both of which we ought to just accept. My experience is that people end up doing the equivalent of saying, "I'm taking my reference class and going home."

A final problem I have with many cases of "reference class forecasting" is that—in addition to everyone coming up with a different reference class—their final answers often seem more specific than I think our state of knowledge should allow. I don't think you *should* be able to tell me that the next major growth mode will have a doubling time of between a month and a year. The alleged outside viewer claims to know too much, once they stake their all on a single preferred reference class. But then what I have just said is an argument for enforced humility—"I don't know, so you can't know either!"—and is automatically suspect on those grounds.

It must be fully conceded and advised that complicated models are hard to fit to limited data, and that when postulating curves which are hard to observe directly or nail down with precision, there is a great deal of room for things to go wrong. It does not follow that "reference class forecasting" is a good solution, or even the merely best solution.

## 3. Some Defenses of a Model of Hard Takeoff

If only for reasons of concreteness, it seems appropriate to summarize my own stance on the intelligence explosion, not just abstractly discuss how to formalize such stances in general.[36] In very concrete terms—leaving out all the abstract principles, microfounda-

36. Peter Cheeseman once told me an anecdote about a speaker at a robotics conference who worked on the more theoretical side of academia, lecturing to an audience of nuts-and-bolts engineers. The talk

tions, and the fundamental question of "What do you think you know and how do you think you know it?"—a "typical" intelligence explosion event as envisioned by Eliezer Yudkowsky might run something like this:

Some sort of AI project run by a hedge fund, academia, Google,[37] or a government, advances to a sufficiently developed level (see section 3.10) that it starts a string of self-improvements that is sustained and does not level off. This cascade of self-improvements might start due to a basic breakthrough by the researchers which enables the AI to understand and redesign more of its own cognitive algorithms. Or a soup of self-modifying systems governed by a fitness evaluator, after undergoing some smaller cascades of self-improvements, might finally begin a cascade which does not level off. Or somebody with money might throw an unprecedented amount of computing power at AI algorithms which don't entirely fail to scale.

Once this AI started on a sustained path of intelligence explosion, there would follow some period of time while the AI was actively self-improving, and perhaps obtaining additional resources, but hadn't yet reached a cognitive level worthy of being called "superintelligence." This time period might be months or years,[38] or days or seconds.[39] I am greatly uncertain of what signs of competence the AI might give over this time, or how its builders or other parties might react to this; but for purposes of intelligence explosion microeconomics, we should temporarily factor out these questions and assume the AI's growth is not being deliberately impeded by any particular agency.

At some point the AI would reach the point where it could solve the protein structure prediction problem and build nanotechnology—or figure out how to control atomic-

---

revolved entirely around equations consisting of upper-case Greek letters. During the Q&A, somebody politely asked the speaker if he could give a concrete example. The speaker thought for a moment and wrote a new set of equations, only this time all the Greek letters were in lowercase.

I try not to be that guy.

37. Larry Page has publicly said that he is specifically interested in "real AI" (Artificial General Intelligence), and some of the researchers in the field are funded by Google. So far as I know, this is still at the level of blue-sky work on basic algorithms and not an attempt to birth The Google in the next five years, but it still seems worth mentioning Google specifically.

38. Any particular AI's characteristic growth path might require centuries to superintelligence—this could conceivably be true even of some modern AIs which are not showing impressive progress—but such AIs end up being irrelevant; some other project which starts later will reach superintelligence first. Unless all AI development pathways require centuries, the surrounding civilization will continue flipping through the deck of AI development projects until it turns up a faster-developing AI.

39. Considering that current CPUs operate at serial speeds of billions of operations per second and that human neurons require at least a millisecond to recover from firing a spike, seconds are potentially long stretches of time for machine intelligences—a second has great serial depth, allowing many causal events to happen in sequence. See section 3.3.

force microscopes to create new tool tips that could be used to build small nanostructures which could build more nanostructures—or perhaps follow some smarter and faster route to rapid infrastructure. An AI that goes past this point can be considered to have reached a threshold of great material capability. From this would probably follow cognitive superintelligence (if not already present); vast computing resources could be quickly accessed to further scale cognitive algorithms.

The further growth trajectory beyond molecular nanotechnology seems mostly irrelevant to present-day policy. An AI with molecular nanotechnology would have sufficient technological advantage, sufficient independence, and sufficient cognitive speed relative to humans that what happened afterward would depend primarily on the AI's preferences. We can try to affect those preferences by wise choice of AI design. But that leads into an entirely different discussion (as remarked on in 1.3), and this latter discussion doesn't seem to depend much on the question of exactly how powerful a superintelligence would become in scenarios where it was already more powerful than the rest of the world economy.

What sort of general beliefs does this concrete scenario of "hard takeoff" imply about returns on cognitive reinvestment?

It supposes that:

- An AI can get major gains rather than minor gains by doing better computer science than its human inventors.
- More generally, it's being supposed that an AI can achieve large gains through better use of computing power it already has, or using only processing power it can rent or otherwise obtain on short timescales—in particular, without setting up new chip factories or doing anything else which would involve a long, unavoidable delay.[40]
- An AI can continue reinvesting these gains until it has a huge cognitive problem-solving advantage over humans.

40. Given a choice of investments, a rational agency will choose the investment with the highest interest rate—the greatest multiplicative factor per unit time. In a context where gains can be *repeatedly reinvested*, an investment that returns 100-fold in one year is vastly inferior to an investment which returns 1.001-fold in one hour. At some point an AI's internal code changes will hit a ceiling, but there's a huge incentive to climb toward, e.g., the protein-structure-prediction threshold by improving code rather than by building chip factories. Buying more CPU time is an intermediate case, but keep in mind that adding hardware also increases the returns on algorithmic improvements (see section 3.1). (This is another reason why I go to some lengths to dissociate my beliefs from any reliance on Moore's Law continuing into the near or distant future. Waiting years for the next generation of chips should not be a preferred modality for an intelligence explosion in progress.)

- This cognitive superintelligence can echo back to tremendous real-world capabilities by solving the protein folding problem, or doing something else even more clever (see section 3.11), starting from the then-existing human technological base.

Even more abstractly, this says that AI self-improvement can operate with $k \gg 1$ and a fast timescale of reinvestment: "prompt supercritical."

But why believe that?

(A question like this is conversationally difficult to answer since different people may think that different parts of the scenario sound most questionable. Also, although I think there is a simple idea at the core, when people ask probing questions the resulting conversations are often much more complicated.[41] Please forgive my answer if it doesn't immediately address the questions at the top of your own priority list; different people have different lists.)

I would start out by saying that the evolutionary history of hominid intelligence doesn't show any signs of diminishing returns—there's no sign that evolution took ten times as long to produce each successive marginal improvement of hominid brains. (Yes, this is hard to quantify, but even so, the anthropological record doesn't look like it should look if there were significantly diminishing returns. See section 3.6.) We have a fairly good mathematical grasp on the processes of evolution and we can well approximate some of the optimization pressures involved; we can say with authority that, in a number of important senses, evolution is extremely inefficient (Yudkowsky 2007). And yet evolution was able to get significant cognitive returns on point mutations, random recombination, and non-foresightful hill climbing of genetically encoded brain architectures. Furthermore, the character of evolution as an optimization process was essentially constant over the course of mammalian evolution—there were no truly fundamental innovations, like the evolutionary invention of sex and sexual recombination, over the relevant timespan.

So if a steady pressure from natural selection realized significant fitness returns from optimizing the intelligence of hominids, then researchers getting smarter at optimizing *themselves* ought to go FOOM.

The "fully naive" argument from Moore's Law folded in on itself asks, "If computing power is doubling every eighteen months, what happens when computers are doing the research?" I don't think this scenario is actually important in practice, mostly because I

41. "The basic idea is simple, but refuting objections can require much more complicated conversations" is not an alarming state of affairs with respect to Occam's Razor; it is common even for correct theories. For example, the core idea of natural selection was much simpler than the conversations that were required to refute simple-sounding objections to it. The added conversational complexity is often carried in by invisible presuppositions of the objection.

expect returns on cognitive algorithms to dominate returns on speed. (The dominant species on the planet is not the one that evolved the fastest neurons.) Nonetheless, if the difficulty curve of Moore's Law was such that humans could climb it at a steady pace, then *accelerating* researchers, researchers whose speed was itself tied to Moore's Law, should arguably be expected to (from our perspective) go FOOM.

The returns on pure speed might be comparatively smaller—sped-up humans would not constitute superintelligences. (For more on returns on pure speed, see section 3.3.) However, faster minds are easier to imagine than smarter minds, and that makes the "folded-in Moore's Law" a simpler illustration of the general idea of folding-in.

Natural selection seems to have climbed a linear or moderately superlinear growth curve of cumulative optimization pressure in versus intelligence out. To "fold in" this curve we consider a scenario where the inherent difficulty of the problem is as before, but instead of minds being improved from the outside by a steady pressure of natural selection, the current optimization power of a mind is determining the speed at which the curve of "cumulative optimization power in" is being traversed. Given the previously described characteristics of the non-folded-in curve, any particular self-improving agency, without outside help, should either bottleneck in the lower parts of the curve (if it is not smart enough to make improvements that are significant compared to those of long-term cumulative evolution), or else go FOOM (if its initial intelligence is sufficiently high to start climbing) and then climb even faster.

We should see a "bottleneck or breakthrough" dichotomy: Any particular self-improving mind either "bottlenecks" without outside help, like all current AIs, or "breaks through" into a fast intelligence explosion.[42] There would be a border between these alternatives containing minds which are seemingly making steady, slow, significant progress at self-improvement; but this border need not be wide, and any such mind would be steadily moving toward the FOOM region of the curve. See section 3.10.

Some amount of my confidence in "AI go FOOM" scenarios also comes from cognitive science (e.g., the study of heuristics and biases) suggesting that humans are, in practice, very far short of optimal design. The broad state of cognitive psychology suggests that "Most humans cannot multiply two three-digit numbers in their heads" is not an unfair indictment—we really are that poorly designed along many dimensions.[43]

42. At least the first part of this prediction seems to be coming true.

43. This is admittedly an impression one picks up from long acquaintance with the field. There is no one single study that conveys, or properly should convey, a strong conclusion that the human mind design is incredibly bad along multiple dimensions. There are representative single examples, like a mind with $10^{14}$ processing elements failing to solve the abstract Wason selection task on the first try. But unless you know the longer story behind that, and how many other results are similar, it doesn't have the same impact.

On a higher level of abstraction, this is saying that there exists great visible headroom for improvement over the human level of intelligence. It's extraordinary that humans manage to play chess using visual recognition systems which evolved to distinguish tigers on the savanna; amazing that we can use brains which evolved to make bows and arrows to program computers; and downright incredible that we can invent new computer science and new cognitive algorithms using brains mostly adapted to modeling and outwitting other humans. But by the standards of computer-based minds that can redesign themselves as required and run error-free algorithms with a billion steps of serial depth, we probably aren't thinking very *efficiently*. (See section 3.5.)

Thus we have specific reason to suspect that cognitive algorithms can be improved beyond the human level—that human brain algorithms aren't any closer to optimal software than human neurons are close to the physical limits of hardware. Even without the embarrassing news from experimental psychology, we could still observe that the inherent difficulty curve for building intelligences has no known reason to possess the specific irregularity of curving sharply upward just after accessing human equivalence. But we also have specific reason to suspect that mind designs can be substantially improved beyond the human level.

That is a rough summary of what I consider the core idea behind my belief that returns on cognitive reinvestments are probably large. You could call this summary the "naive" view of returns on improving cognitive algorithms, by analogy with the naive theory of how to fold in Moore's Law. We can drill down and ask more sophisticated questions, but it's worth remembering that when done correctly, more sophisticated analysis quite often says that the naive answer is right. Somebody who'd never studied General Relativity as a formal theory of gravitation might naively expect that jumping off a tall cliff would make you fall down and go splat; and in this case it turns out that the sophisticated prediction agrees with the naive one.

Thus, keeping in mind that we are not obligated to arrive at any impressively nonobvious "conclusions," let us consider some nonobvious subtleties of argument.

In the next subsections we will consider:

1. What the fossil record actually tells us about returns on brain size, given that most of the difference between *Homo sapiens* and *Australopithecus* was probably improved algorithms.

2. How to divide credit for the human-chimpanzee performance gap between "humans are individually smarter than chimpanzees" and "the hominid transition involved a one-time qualitative gain from being able to accumulate knowledge." More generally, the problem of how to analyze supposed *one-time gains* that should allegedly be factored out of predicted future growth.

3. How returns on speed (serial causal depth) contrast with returns from parallelism; how faster thought seems to contrast with more thought. Whether sensing and manipulating technologies are likely to present a bottleneck for faster thinkers, and if so, how large a bottleneck.

4. How human populations seem to scale in problem-solving power; some reasons to believe that we scale more inefficiently than machine intelligences would. Garry Kasparov's chess match versus The World, which Kasparov won.

5. Some inefficiencies that might accumulate in an estimate of humanity's net computational efficiency on a cognitive problem.

6. What the anthropological record actually tells us about cognitive returns on cumulative selection pressure, given that selection pressures were probably increasing over the course of hominid history. How observed history would be expected to look different if there were diminishing returns on cognition or evolution.

7. How to relate the curves for evolutionary difficulty, human-engineering difficulty, and AI-engineering difficulty, considering that they are almost certainly different.

8. Correcting for *anthropic bias* in trying to estimate the intrinsic "difficulty" of hominid-level intelligence from observing that intelligence evolved here on Earth. (The problem being that on planets where intelligence does not evolve, there is no one to observe its absence.)

9. The question of whether to expect a "local" (one-project) or "global" (whole economy) FOOM, and how quantitative returns on cognitive reinvestment interact with that.

10. The great open uncertainty about the minimal conditions for starting a FOOM; why I. J. Good's original postulate of starting from "ultraintelligence" seems much too strong (sufficient, but very far above what is necessary).

11. The enhanced importance of unknown unknowns in intelligence explosion scenarios, since a smarter-than-human intelligence will selectively seek out and exploit useful possibilities implied by flaws or gaps in our current knowledge.

I would finally remark that going into depth on the pro-FOOM stance should not operate to prejudice the reader in favor of other stances. Defending only one stance at great length may make it look like a huge edifice of argument that could potentially topple, whereas other viewpoints such as "A collective of interacting AIs will have $k \approx 1^+$ and grow at a manageable, human-like exponential pace, just like the world economy" may sound "simpler" because their points and counterpoints have not yet been explored.

But of course (so far as the author believes) such other outcomes would be even harder to defend in depth.[44] Every argument for the intelligence explosion is, when negated, an argument for an intelligence nonexplosion. To the extent the *negation* of each argument here might sound less than perfectly plausible, other possible outcomes would not sound any *more* plausible when argued to this depth of point and counterpoint.

### 3.1. Returns on Brain Size

Many cases where we'd like to reason from historical returns on cognitive investment are complicated by unfortunately narrow data. All the most impressive cognitive returns are from a single species, namely *Homo sapiens*.

Humans have brains around four times the size of chimpanzees' . . . but this tells us very little because most of the differences between humans and chimps are almost certainly algorithmic. If just taking an *Australopithecus* brain and scaling it up by a factor of four produced a human, the evolutionary road from *Australopithecus* to *Homo sapiens* would probably have been much shorter; simple factors like the size of an organ can change quickly in the face of strong evolutionary pressures.

Based on historical observation, we can say with authority that going from *Australopithecus* to *Homo sapiens* did not in fact require a hundredfold increase in brain size *plus* improved algorithms—we can refute the assertion that even after taking into account five million years of evolving better cognitive algorithms, a hundredfold increase in hardware was required to accommodate the new algorithms. This may not sound like much, but it does argue against models which block an intelligence explosion by always requiring exponentially increasing hardware for linear cognitive gains.[45]

A nonobvious further implication of observed history is that improvements in cognitive algorithms along the way to *Homo sapiens* must have increased rather than decreased

---

44. Robin Hanson has defended the "global exponential economic speedup" thesis at moderate length, in the Yudkowsky-Hanson AI-Foom debate and in several papers, and the reader is invited to explore these.

I am not aware of anyone who has defended an "intelligence fizzle" seriously and at great length, but this of course may reflect a selection effect. If you believe nothing interesting will happen, you don't believe there's anything worth writing a paper on.

45. I'm pretty sure I've heard this argued several times, but unfortunately I neglected to save the references; please contribute a reference if you've got one. Obviously, the speakers I remember were using this argument to confidently dismiss the possibility of superhuman machine intelligence, and it did not occur to them that the same argument might also apply to the hominid anthropological record.

If this seems so silly that you doubt anyone really believes it, consider that "the intelligence explosion is impossible because Turing machines can't promote themselves to hypercomputers" is worse, and see Bringsjord (2012) for the appropriate citation by a distinguished scientist.

We can be reasonably extremely confident that human intelligence does not take advantage of quantum computation (Tegmark 2000). The computing elements of the brain are too large and too hot.

the marginal fitness returns on larger brains and further-increased intelligence, because the new equilibrium brain size was four times as large.

To elaborate on this reasoning: A rational agency will invest such that the marginal returns on all its fungible investments are approximately equal. If investment X were yielding more on the margins than investment Y, it would make sense to divert resources from Y to X. But then diminishing returns would reduce the yield on further investments in X and increase the yield on further investments in Y; so after shifting some resources from Y to X, a new equilibrium would be found in which the marginal returns on investments were again approximately equal.

Thus we can reasonably expect that for any species in a rough evolutionary equilibrium, each marginal added unit of ATP (roughly, metabolic energy) will yield around the same increment of inclusive fitness whether it is invested in the organism's immune system or in its brain. If it were systematically true that adding one marginal unit of ATP yielded much higher returns in the immune system compared to the brain, that species would experience a strong selection pressure in favor of diverting ATP from organisms' brains to their immune systems. Evolution measures all its returns in the common currency of inclusive genetic fitness, and ATP is a fungible resource that can easily be spent anywhere in the body.

The human brain consumes roughly 20% of the ATP used in the human body, an enormous metabolic investment. Suppose a positive mutation makes it possible to accomplish the same cognitive work using only 19% of the body's ATP—with this new, more efficient neural algorithm, the same cognitive work can be done by a smaller brain. If we are in a regime of strongly diminishing fitness returns on cognition[46] *or* strongly diminishing cognitive returns on adding further neurons,[47] then we should expect the

46. Suppose your rooms are already lit as brightly as you like, and then someone offers you cheaper, more energy-efficient light bulbs. You will light your room at the same brightness as before and decrease your total spending on lighting. Similarly, if you are already thinking well enough to outwit the average deer, and adding more brains does not let you outwit deer any better because you are already smarter than a deer (diminishing fitness returns on further cognition), then evolving more efficient brain algorithms will lead to evolving a smaller brain that does the same work.

47. Suppose that every meal requires a hot dog and a bun; that it takes 1 unit of effort to produce each bun; and that each successive hot dog requires 1 more unit of labor to produce, starting from 1 unit for the first hot dog. Thus it takes 6 units to produce 3 hot dogs and 45 units to produce 9 hot dogs. Suppose we're currently eating 9 meals based on $45 + 9 = 54$ total units of effort. Then even a magical bun factory which eliminates all of the labor in producing buns will not enable the production of 10 meals, due to the increasing cost of hot dogs. Similarly if we can recover large gains by improving the efficiency of one part of the brain, but the limiting factor is another brain part that scales very poorly, then the fact that we improved a brain algorithm well enough to significantly shrink the total cost of the brain doesn't necessarily mean that we're in a regime where we can do significantly more total cognition by reinvesting the saved neurons.

brain to shrink as the result of this innovation, doing the same total work at a lower price. But in observed history, hominid brains grew larger instead, paying a greater metabolic price to do even more cognitive work. It follows that over the course of hominid evolution there were both significant marginal fitness returns on improved cognition *and* significant marginal cognitive returns on larger brains.

In economics this is known as the Jevons paradox—the counterintuitive result that making lighting more electrically efficient or making electricity cheaper can increase the total money spent on lighting. The returns on buying lighting go up, so people buy more of it and the total expenditure increases. Similarly, some of the improvements to hominid brain algorithms over the course of hominid evolution must have increased the marginal fitness returns of spending even more ATP on the brain. The equilibrium size of the brain, and its total resource cost, shifted upward as cognitive algorithms improved.

Since human brains are around four times the size of chimpanzee brains, we can conclude that our increased efficiency (cognitive yield on fungible biological resources) increased the marginal returns on brains such that the new equilibrium brain size was around four times as large. This unfortunately tells us very little quantitatively about the return-on-investment curves for larger brains and constant algorithms—just the qualitative truths that the improved algorithms did increase marginal cognitive returns on brain size, and that there weren't sharply diminishing returns on fitness from doing increased amounts of cognitive labor.

It's not clear to me how much we should conclude from brain sizes increasing by a factor of *only* four—whether we can upper-bound the returns on hardware this way. As I understand it, human-sized heads lead to difficult childbirth due to difficulties of the baby's head passing the birth canal. This is an adequate explanation for why we wouldn't see superintelligent mutants with triple-sized heads, even if triple-sized heads could yield superintelligence. On the other hand, it's not clear that human head sizes are *hard* up against this sort of wall—some people have above-average-sized heads without their mothers being dead. Furthermore, Neanderthals may have had larger brains than modern humans (Ponce de León et al. 2008).[48] So we are probably licensed to conclude that there has not been a strong selection pressure for larger brains, as such, over very recent evolutionary history.[49]

48. Neanderthals were not our direct ancestors (although some interbreeding may have occurred), but they were sufficiently closely related that their larger cranial capacities are relevant evidence.

49. It is plausible that the marginal fitness returns on cognition have leveled off sharply enough that improvements in cognitive efficiency have shifted the total resource cost of brains downward rather than upward over very recent history. If true, this is not the same as *Homo sapiens sapiens* becoming stupider or even staying the same intelligence. But it does imply that either marginal fitness returns on cognition or

There are two steps in the derivation of a fitness return from increased brain size: a cognitive return on brain size and a fitness return on cognition. For example, John von Neumann[50] had only one child, so the transmission of cognitive returns to fitness returns might not be perfectly efficient. We can upper bound the fitness returns on larger brains by observing that *Homo sapiens* are not hard up against the wall of head size and that Neanderthals may have had even larger brains. This doesn't say how much of that bound on returns is about fitness returns on cognition versus cognitive returns on brain size.

Do variations in brain size within *Homo sapiens* let us conclude much about cognitive returns? Variance in brain size correlates around 0.3 with variance in measured IQ, but there are many plausible confounders such as childhood nutrition or childhood resistance to parasites. The best we can say is that John von Neumann did not seem to require a brain exponentially larger than that of an average human, or even twice as large as that of an average human, while displaying scientific productivity well in excess of twice that of an average human being of his era. But this presumably isn't telling us about enormous returns from small increases in brain size; it's much more likely telling us that other factors can produce great increases in scientific productivity without requiring large increases in brain size. We can also say that it's not possible that a 25% larger brain automatically yields superintelligence, because that's within the range of existing variance.

The main lesson I end up deriving is that intelligence improvement has not *required* exponential increases in computing power, and that marginal fitness returns on increased brain sizes were significant over the course of hominid evolution. This corresponds to AI growth models in which large cognitive gains by the AI can be accommodated by acquiring already-built computing resources, without needing to build new basic chip technologies.

Just as an improved algorithm can increase the marginal returns on adding further hardware (because it is running a better algorithm), additional hardware can increase the marginal returns on improved cognitive algorithms (because they are running on more hardware).[51] In everyday life, we usually expect feedback loops of this sort to die down, but in the case of hominid evolution there was in fact strong continued growth, so it's

---

marginal cognitive returns on brain scaling have leveled off significantly compared to earlier evolutionary history.

50. I often use John von Neumann to exemplify the far end of the human intelligence distribution, because he is widely reputed to have been the smartest human being who ever lived and all the other great geniuses of his era were scared of him. Hans Bethe said of him, "I have sometimes wondered whether a brain like von Neumann's does not indicate a species superior to that of man" (Blair 1957).

51. Purchasing a $1,000,000 innovation that improves all your processes by 1% is a terrible investment for a $10,000,000 company and a great investment for a $1,000,000,000 company.

possible that a feedback loop of this sort played a significant role. Analogously it may be possible for an AI design to go FOOM just by adding vastly more computing power, the way a nuclear pile goes critical just by adding more identical uranium bricks; the added hardware could multiply the returns on all cognitive investments, and this could send the system from $k < 1$ to $k > 1$. Unfortunately, I see very little way to get any sort of quantitative grasp on this probability, apart from noting the qualitative possibility.[52]

In general, increased "size" is a kind of cognitive investment about which I think I know relatively little. In AI it is usual for hardware improvements to contribute lower gains than software improvements—with improved hardware still being critical, because with a sufficiently weak computer, the initial algorithms can perform so poorly that it doesn't pay incrementally to improve them.[53] Even so, most of the story in AI has always been about software rather than hardware, and with hominid brain sizes increasing by a mere factor of four over five million years, this seems to have been true for hominid evolution as well.

Attempts to predict the advent of AI by graphing Moore's Law and considering the mere addition of computing power appear entirely pointless to me given this overall state of knowledge. The cognitive returns on hardware are always changing as a function of improved algorithms; there is no calculable constant threshold to be crossed.

### 3.2. One-Time Gains

On an intuitive level, it seems obvious that the human species has accumulated cognitive returns sufficiently in excess of the chimpanzee species; we landed on the Moon and they didn't. Trying to get a quantitative grasp on the "cognitive returns on humans," and how much they actually exceed the cognitive returns on chimpanzees, is greatly complicated by the following facts:

- There are many more humans than chimpanzees.

52. This scenario is not to be confused with a large supercomputer spontaneously developing consciousness, which Pat Cadigan accurately observed to be analogous to the old theory that dirty shirts and straw would spontaneously generate mice. Rather, the concern here is that you already have an AI design which is qualitatively capable of significant self-improvement, and it goes critical after some incautious group with lots of computing resources gets excited about those wonderful early results and tries running the AI on a hundred thousand times as much computing power.

53. If hominids were limited to spider-sized brains, it would be much harder to develop human-level intelligence, because the incremental fitness returns on improved algorithms would be lower (since each algorithm runs on less hardware). In general, a positive mutation that conveys half as much advantage takes twice as long to rise to fixation, and has half the chance of doing so at all. So if you diminish the fitness returns to each step along an adaptive pathway by three orders of magnitude, the evolutionary outcome is not "this adaptation takes longer to evolve" but "this adaptation does not evolve at all."

- Humans can communicate with each other much better than chimpanzees.

This implies the possibility that cognitive returns on improved brain algorithms (for humans vs. chimpanzees) might be smaller than the moon landing would suggest. Cognitive returns from *better-cumulating* optimization, by a much more *numerous* species that can use language to convey knowledge across brains, should not be confused with any inherent power of a single human brain. We know that humans have nuclear weapons and chimpanzees don't. But to the extent we attribute this to larger human populations, we must not be attributing it to humans having writing; and to the extent we attribute it to humans having writing, we must not be attributing it to humans having larger brains and improved cognitive algorithms.[54]

"That's silly," you reply. "Obviously you need writing *and* human general intelligence before you can invent science and have technology accumulate to the level of nuclear weapons. Even if chimpanzees had some way to pass on the knowledge they possessed and do cumulative thinking—say, if you used brain-computer interfaces to directly transfer skills from one chimpanzee to another—they'd probably still never understand linear algebra, even in a million years. It's not a question of communication versus individual intelligence, there's a joint causal dependency."

Even so (goes the counter-counterpoint) it remains obvious that discovering and using electricity is not a pure property of a single human brain. Speech and writing, as inventions enabled by hominid intelligence, induce a change in the character of cognitive intelligence as an optimization process: thinking time cumulates more strongly across populations and centuries. To the extent that we're skeptical that any further innovations of this sort exist, we might expect the grand returns of human intelligence to be a mostly one-time affair, rather than a repeatable event that scales proportionally with larger brains or further-improved cognitive algorithms. If being able to cumulate knowledge is an absolute threshold which has already been crossed, we can't expect to see repeatable cognitive returns from crossing it again and again.

But then (says the counter-counter-counterpoint) we may not be all the way across the communication threshold. Suppose humans could not only talk to each other but perfectly transfer complete cognitive skills, and could not only reproduce humans in general but duplicate thousands of mutually telepathic Einsteins, the way AIs could copy themselves and transfer thoughts. Even if communication is a one-time threshold, we could be more like 1% over the threshold than 99% over it.

54. Suppose I know that your investment portfolio returned 20% last year. The higher the return of the stocks in your portfolio, the less I must expect the bonds in your portfolio to have returned, and vice versa.

However (replies the counter[4]-point) if the ability to cumulate knowledge is still qualitatively present among humans, doing so more efficiently might not yield marginal returns proportional to crossing the initial threshold. Suppose there's a constant population of a hundred million people, and returns to the civilization are determined by the most cumulated cognitive labor. Going from 0% cumulation to 1% cumulation between entities might multiply total returns much more than the further multiplicative factor in going from 1% cumulation to 99% cumulation. In this scenario, a thousand 1%-cumulant entities can outcompete a hundred million 0%-cumulant entities, and yet a thousand perfectly cumulant entities cannot outcompete a hundred million 1% cumulant entities, depending on the details of your assumptions.

A counter[5]-point is that this would not be a good model of piles of uranium bricks with neutron-absorbing impurities; any degree of noise or inefficiency would interfere with the clarity of the above conclusion. A further counter[5]-point is to ask about the invention of the printing press and the subsequent industrial revolution—if the one-time threshold model is true, why did the printing press enable civilizational returns that seemed to be well above those of writing or speech?

A different one-time threshold that spawns a similar line of argument revolves around human generality—the way that we can grasp some concepts that chimpanzees can't represent at all, like the number thirty-seven. The science-fiction novel *Schild's Ladder*, by Greg Egan (2002), supposes a "General Intelligence Theorem" to the effect that once you get to the human level, you're done—you can think about anything thinkable. Hence there are no further gains from further generality; and that was why, in Egan's depicted future, there were no superintelligences despite all the human-level minds running on fast computers.

The obvious inspiration for a "General Intelligence Theorem" is the Church-Turing Thesis: Any computer that can simulate a universal Turing machine is capable of simulating any member of a very large class of systems, which class seems to include the laws of physics and hence everything in the real universe. Once you show you can encode a single universal Turing machine in Conway's Game of Life, then the Game of Life is said to be "Turing complete" because we can encode any other Turing machine inside the universal machine we already built.

The argument for a one-time threshold of generality seems to me much weaker than the argument from communication. Many humans have tried and failed to understand linear algebra. Some humans (however unjust this feature of our world may be) probably cannot understand linear algebra, period.[55] Such humans could, in principle, if immortal

55. Until technology advances to the point of direct cognitive enhancement of humans. I don't believe in giving up when it comes to this sort of thing.

and never bored, take an infinitely long piece of paper tape and simulate by hand a giant Turing machine simulating John von Neumann. But they still wouldn't understand linear algebra; their own brains, as opposed to the paper tape, would not contain any representations apt for manipulating linear algebra.[56] So being over the Church-Turing threshold does not imply a brain with apt native representations for manipulating every possible sort of concept. An immortal mouse would also be over this threshold—most complex systems are—while still experiencing lesser cognitive returns than humans over the timescales of interest. There is also visible headroom above the human level; an obvious future threshold of cognitive generality is the ability to manipulate your source code so as to compose new underlying cognitive representations for any problem you encounter. If a true threshold of cognitive generality exists—if there is any sort of mind that can quickly give itself apt representations for almost any sort of solvable problem—we are under that threshold, not over it. I usually say that what distinguishes humans from chimpanzees is "significantly more generally applicable intelligence" rather than "general intelligence." One could perhaps count humans as being one percent over a threshold of what can possibly be thought about; but relative to the case of communication, it seems much harder to write out an argument that being one percent over the threshold of generality offers most of the marginal returns.

The main plausible source of such an argument would be an "end of science" scenario in which most of the interesting, exploitable possibilities offered by the physical universe could all be understood by some threshold level of generality, and thus there would be no significant returns to generality beyond this point. Humans have not developed many technologies that seem foreseeable in some sense (e.g., we do not yet have molecular nanotechnology) but, amazingly enough, all of the future technologies we can imagine from our current level seem to be graspable using human-level abilities for abstraction. This, however, is not strong evidence that no greater capacity for abstraction can be helpful in realizing all important technological possibilities.

In sum, and taking into account all three of the arguments listed above, we get a combined argument as follows:

The Big Marginal Return on humans over chimpanzees is mostly about *large numbers* of humans, *sharing knowledge* above a sharp *threshold of abstraction*, being more impressive than the sort of thinking that can be done by *one* chimpanzee who cannot communicate with other chimps and is qualitatively incapable of grasping algebra. Then since very little of the Big Marginal Return was really about improving cognitive algorithms or increasing brain sizes apart from that, we have no reason to believe that there were any

56. Note the resemblance to the standard reply (Cole 2013) to Searle's Chinese Room argument.

repeatable gains of this sort. Most of the chimp-human difference is from cumulating total power rather than individual humans being smarter; you can't get human-versus-chimp gains just from having a larger brain than one human. To the extent humans are qualitatively smarter than chimps, it's because we crossed a qualitative threshold which lets (unusually smart) humans learn linear algebra. But now that some of us can learn linear algebra, there are no more thresholds like that. When all of this is taken into account, it explains away most of the human bonanza and doesn't leave much to be attributed just to evolution optimizing cognitive algorithms *qua* algorithms and hominid brain sizes increasing by a factor of four. So we have no reason to suppose that bigger brains or better algorithms could allow an AI to experience the same sort of increased cognitive returns above humans as humans have above chimps.

The above argument postulates one-time gains which all lie in our past, with no similar gains in the future. In a sense, all gains from optimization are one-time—you cannot invent the steam engine twice, or repeat the same positive mutation—and yet to expect this ongoing stream of one-time gains to halt at any particular point seems unjustified. In general, postulated one-time gains—whether from a single threshold of communication, a single threshold of generality/abstraction, etc.—seem hard to falsify or confirm by staring at raw growth records. In general, my reply is that I'm quite willing to believe that hominids have crossed qualitative thresholds, less willing to believe that such a young species as ours is already 99% over a threshold rather than 10% or 0.03% over that threshold, and extremely skeptical that all the big thresholds are already in our past and none lie in our future. Especially when humans seem to lack all sorts of neat features such as the ability to expand indefinitely onto new hardware, the ability to rewrite our own source code, the ability to run error-free cognitive processes of great serial depth, etc.[57]

It is certainly a feature of the design landscape that it contains large one-time gains—significant thresholds that can only be crossed once. It is less plausible that hominid evolution crossed them *all* and arrived at the qualitative limits of mind—especially when many plausible further thresholds seem clearly visible even from here.

### 3.3. Returns on Speed

By the standards of the eleventh century, the early twenty-first century can do things that would seem like "magic" in the sense that nobody in the eleventh century imagined them, let alone concluded that they would be possible.[58] What separates the early twenty-first century from the eleventh?

57. Not to mention everything that the human author hasn't even thought of yet. See section 3.11.

58. See again section 3.11.

Gregory Clark (2007) has suggested, based on demographic data from British merchants and shopkeepers, that more conscientious individuals were having better financial success and more children, and to the extent that conscientiousness is hereditary this would necessarily imply natural selection; thus Clark has argued that there was probably some degree of genetic change supporting the Industrial Revolution.

But this seems like only a small caveat to the far more obvious explanation that what separated the eleventh and twenty-first centuries was time.

What is time? Leaving aside some interesting but not overwhelmingly relevant answers from fundamental physics,[59] when considered as an economic resource, "time" is the ability for events to happen one after another. You cannot invent jet planes at the same time as internal combustion engines; to invent transistors, somebody must have already finished discovering electricity and told you about it. The twenty-first century is separated from the eleventh century by a series of discoveries and technological developments that did in fact occur one after another and would have been significantly more difficult to do in parallel.

A more descriptive name for this quality than "time" might be "serial causal depth." The saying in software industry goes, "Nine women can't birth a baby in one month," indicating that you can't just add more people to speed up a project; a project requires time, sequential hours, as opposed to just a total number of human-hours of labor. Intel has not hired twice as many researchers as its current number and produced new generations of chips twice as fast.[60] This implies that Intel thinks its largest future returns will come from discoveries that must be made after current discoveries (as opposed to most future returns coming from discoveries that can all be reached by one step in a flat search space and hence could be reached twice as fast by twice as many researchers).[61]

Similarly, the "hundred-step rule" in neuroscience (Feldman and Ballard 1982) says that since human neurons can only fire around one hundred times per second, any computational process that humans seem to do in real time must take at most one hundred *serial* steps—that is, one hundred steps that must happen one after another.

59. See, e.g., Barbour (1999).

60. With Intel's R&D cost around 17% of its sales, this wouldn't be easy, but it would be possible.

61. If Intel thought that its current researchers would exhaust the entire search space, or exhaust all marginally valuable low-hanging fruits in a flat search space, then Intel would be making plans to terminate or scale down its R&D spending after one more generation. Doing research with a certain amount of parallelism that is neither the maximum or minimum you could possibly manage implies an expected equilibrium, relative to your present and future returns on technology, of how many fruits you can find at the immediate next level of the search space, versus the improved returns on searching later after you can build on previous discoveries. (Carl Shulman commented on a draft of this paper that Intel may also rationally wait because it expects to build on discoveries made outside Intel.)

There are billions of neurons in the visual cortex and so it is reasonable to suppose a visual process that involves billions of computational steps. But you cannot suppose that A happens, and that B which depends on A happens, and that C which depends on B happens, and so on for a billion steps. You cannot have a series of events like that inside a human brain; the series of events is too causally deep, and the human brain is too serially shallow. You can't even have a million-step serial process inside a modern-day factory; it would take far too long and be far too expensive to manufacture anything that required a million manufacturing steps to occur one after another. That kind of serial causal depth can *only* occur inside a computer.

This is a great part of what makes computers useful, along with their ability to carry out formal processes exactly: computers contain huge amounts of time, in the sense of containing tremendous serial depths of causal events. Since the Cambrian explosion and the rise of anatomical multicellular organisms $2 \times 10^{11}$ days ago, your line of direct descent might be perhaps $10^8$ or $10^{11}$ generations deep. If humans had spoken continuously to each other since 150,000 years ago, one utterance per five seconds, the longest continuous conversation could have contained $\sim 10^{12}$ statements one after another. A 2013-era CPU running for one day can contain $\sim 10^{14}$ programmable events occurring one after another, or $\sim 10^{16}$ events if you run it for one year.[62] Of course, if we are talking about a six-core CPU, then that is at most six things that could be happening at the same time, and a floating-point multiplication is a rather simple event. Still, when I contemplate statistics like those above, I am struck by a vertiginous sense of what incredibly poor use we make of computers.

Although I used to go around asking, "If Moore's Law says that computing speeds double every eighteen months, what happens when computers are doing the research?"[63] I no longer think that Moore's Law will play much of a role in the intelligence explosion, partially because I expect returns on algorithms to dominate, and partially because I would expect an AI to prefer ways to scale itself onto more existing hardware rather than waiting for a new generation of chips to be produced in Intel-style factories. The latter form of investment has such a slow timescale, and hence such a low interest rate, that I would only expect it to be undertaken if all other self-improvement alternatives had bottlenecked before reaching the point of solving protein structure prediction or otherwise bypassing large human-style factories.

62. Almost the same would be true of a 2008-era CPU, since the Moore's-like law for serial depth has almost completely broken down. Though CPUs are also not getting any slower, and the artifacts we have already created seem rather formidable in an absolute sense.

63. I was then seventeen years old.

Since computers are well known to be fast, it is a very widespread speculation that strong AIs would think very fast because computers would be very fast, and hence that such AIs would rapidly acquire advantages of the sort we associate with older human civilizations, usually improved science and technology.[64] Two objections that have been offered against this idea are (a) that the first sufficiently advanced AI might be very slow while already running on a large fraction of all available computing power, and hence hard to speed up without waiting on Moore's Law,[65] and (b) that fast thinking may prove useless without fast sensors and fast motor manipulators.[66]

Let us consider first the prospect of an advanced AI already running on so much computing power that it is hard to speed up. I find this scenario somewhat hard to analyze because I expect AI to be mostly about algorithms rather than lots of hardware, but I can't rule out scenarios where the AI is developed by some large agency which was running its AI project on huge amounts of hardware from the beginning. This should not make the AI slow in all aspects; any AI with a certain amount of self-reprogramming ability ought to be able to perform many particular kinds of cognition very quickly—to take one extreme example, it shouldn't be slower than humans at arithmetic, even conscious arithmetic. But the AI's overall thought processes might still be slower than human, albeit presumably not so slow that the programmers and researchers are too bored to work effectively on the project or try to train and raise the AI. Thus I cannot say that the overall scenario is implausible. I do note that to the extent that an AI is running on more hardware and has worse algorithms, *ceteris paribus*, you would expect greater gains from improving the algorithms. Trying to deliberately create a slow AI already running on vast amounts of hardware, in hopes of guaranteeing sufficient time to react, may not actually serve to slow down the overall growth curve—it may prove to be the equivalent of starting out the AI with much more hardware than it would have had otherwise, hence greater returns on improving its algorithms. I am generally uncertain about this point.

On the input-output side, there are various Moore's-like curves for sensing and manipulating, but their exponents tend to be lower than the curves for pure computer technologies. If you extrapolated this trend outward without further change, then the pure

64. As the fourth-century Chinese philosopher Xiaoguang Li once observed, we tend to think of earlier civilizations as being more venerable, like a wise old ancestor who has seen many things; but in fact later civilizations are older than earlier civilizations, because the future has a longer history than the past. Thus I hope it will increase, rather than decrease, your opinion of his wisdom if I now inform you that actually Xiaoguang "Mike" Li is a friend of mine who observed this in 2002.

65. This has mostly come up in personal conversation with friends; I'm not sure I've seen a print source.

66. The author is reasonably sure he has seen this objection in print, but failed again to collect the reference at the time.

scenario of "Moore's Law with computer-based researchers" would soon bottleneck on the fast-thinking researchers waiting through their molasses-slow ability to manipulate clumsy robotic hands to perform experiments and actually observe the results.

The field of high-energy physics, for example, seems limited by the expense and delay of constructing particle accelerators. Likewise, subfields of astronomy revolve around expensive space telescopes. These fields seem more sensory-bounded than thinking-bounded, relative to the characteristic intelligence of the researchers. It's possible that sufficiently smarter scientists could get more mileage out of information already gathered, or ask better questions. But at the very least, we can say that there's no humanly-obvious way to speed up high-energy physics with faster-thinking human physicists, and it's easy to imagine that doubling the speed of all the human astronomers, while leaving them otherwise unchanged, would just make them twice as frustrated about telescope time as at present.

At the opposite extreme, theoretical mathematics stands as an example of a field which is limited *only* by the thinking speed of its human researchers (computer assistance currently being a rare exception, rather than the rule). It is interesting to ask whether we should describe progress in mathematics as (1) continuing at mostly the same pace as anything else humans do, or (2) far outstripping progress in every other human endeavor, such that there is no nonmathematical human accomplishment comparable in depth to Andrew Wiles's proof of Fermat's Last Theorem (Wiles 1995).

The main counterpoint to the argument from the slower Moore's-like laws for sensorimotor technologies is that since currently human brains cannot be sped up, and humans are still doing most of the physical labor, there hasn't yet been a strong incentive to produce faster and faster manipulators—slow human brains would still be the limiting factor. But if in the future sensors or manipulators are the limiting factor, most investment by a rational agency will tend to flow toward improving that factor. If slow manipulators are holding everything back, this greatly increases returns on faster manipulators and decreases returns on everything else. But with current technology it is not possible to invest in faster brains for researchers, so it shouldn't be surprising that the speed of researcher thought often is the limiting resource. Any lab that shuts down overnight so its researchers can sleep must be limited by serial cause and effect in researcher brains more than serial cause and effect in instruments—researchers who could work without sleep would correspondingly speed up the lab. In contrast, in astronomy and high-energy physics every minute of apparatus time is scheduled, and shutting down the apparatus overnight would be unthinkable. That most human research labs do cease operation overnight implies that most areas of research are not sensorimotor bounded.

However, rational redistribution of investments to improved sensors and manipulators does not imply that the new resulting equilibrium is one of fast progress. The

counter-counterpoint is that, even so, improved sensors and manipulators are slow to construct compared to just rewriting an algorithm to do cognitive work faster. Hence sensorimotor bandwidth might end up as a limiting factor for an AI going FOOM over short timescales; the problem of constructing new sensors and manipulators might act as metaphorical delayed neutrons that prevent *prompt* criticality. This delay would still exist so long as there were pragmatically real limits on how useful it is to think in the absence of experiential data and the ability to exert power on the world.

A counter-counter-counterpoint is that if, for example, protein structure prediction can be solved as a purely cognitive problem,[67] then molecular nanotechnology is liable to follow very soon thereafter. It is plausible that even a superintelligence might take a while to construct advanced tools if dropped into the thirteenth century with no other knowledge of physics or chemistry.[68] It's less plausible (says the counter-counter-counterargument) that a superintelligence would be similarly bounded in a modern era where protein synthesis and picosecond cameras already exist, and vast amounts of pregathered data are available.[69] Rather than imagining sensorimotor bounding as the equivalent of some poor blind spirit in a locked box, we should imagine an entire human civilization in a locked box, doing the equivalent of cryptography to extract every last iota of inference out of every bit of sensory data, carefully plotting the fastest paths to greater potency using its currently conserved motor bandwidth, using every possible avenue of affecting the world to, as quickly as possible, obtain faster ways of affecting the world. See here for an informal exposition.[70]

I would summarize my views on "speed" or "causal depth" by saying that, contrary to the views of a past Eliezer Yudkowsky separated from my present self by sixteen

67. Note that in some cases the frontier of modern protein structure prediction and protein design is crowdsourced human guessing, e.g., the Foldit project. This suggests that there are gains from applying better cognitive algorithms to protein folding.

68. It's not *certain* that it would take the superintelligence a long time to do anything, because the putative superintelligence is much smarter than you and therefore you cannot exhaustively imagine or search the options it would have available. See section 3.11.

69. Some basic formalisms in computer science suggest fundamentally different learning rates depending on whether you can ask your own questions or only observe the answers to large pools of pre-asked questions. On the other hand, there is also a strong case to be made that humans are overwhelmingly inefficient at constraining probability distributions using the evidence they have already gathered.

70. An intelligence explosion that seems incredibly fast to a human might take place over a long serial depth of parallel efforts, most of which fail, learning from experience, updating strategies, waiting to learn the results of distant experiments, etc., which would appear frustratingly slow to a human who had to perform similar work. Or in implausibly anthropomorphic terms, "Sure, from your perspective it only took me four days to take over the world, but do you have any idea how long that was for *me*? I had to wait twenty thousand subjective years for my custom-ordered proteins to arrive!"

years of "time,"[71] it doesn't seem very probable that returns on hardware speed will be a key ongoing factor in an intelligence explosion. Even Intel constructing new chip factories hasn't increased serial speeds very much since 2004, at least as of 2013. Better algorithms or hardware scaling could decrease the serial burden of a thought and allow more thoughts to occur in serial rather than parallel; it seems extremely plausible that a humanly designed AI will start out with a huge excess burden of serial difficulty, and hence that improving cognitive algorithms or hardware scaling will result in a possibly gradual, possibly one-time huge gain in effective cognitive speed. Cognitive speed outstripping sensorimotor bandwidth in a certain fundamental sense is also very plausible for pre-nanotechnological stages of growth.

The main policy-relevant questions would seem to be:

1. At which stage (if any) of growth will an AI be able to generate new technological capacities of the sort that human civilizations seem to invent "over time," and how quickly?

2. At which stage (if any) of an ongoing intelligence explosion, from which sorts of starting states, will which events being produced by the AI exceed in speed the reactions of (1) human bureaucracies and governments with great power (weeks or months) and (2) individual humans with relatively lesser power (minutes or seconds)?

I would expect that some sort of incredibly fast thinking is likely to arrive at some point, because current CPUs are already very serially fast compared to human brains; what stage of growth corresponds to this is hard to guess. I've also argued that the "high-speed spirit trapped in a statue" visualization is inappropriate, and "high-speed human civilization trapped in a box with slow Internet access" seems like a better way of looking at it. We can visualize some clear-seeming paths from cognitive power to fast infrastructure, like cracking the protein structure prediction problem. I would summarize my view on this question by saying that, although high cognitive speeds may indeed lead to time spent sensorimotor bounded, the total amount of this time may not seem very large from outside—certainly a high-speed human civilization trapped inside a box with Internet access would be trying to graduate to faster manipulators as quickly as possible.

71. Albeit, in accordance with the general theme of embarrassingly overwhelming human inefficiency, the actual thought processes separating $\text{Yudkowsky}_{1997}$ from $\text{Yudkowsky}_{2013}$ would probably work out to twenty days of serially sequenced thoughts or something like that. Maybe much less. Certainly not sixteen years of solid sequential thinking.

### 3.4. Returns on Population

As remarked in section 3.3, the degree to which an AI can be competitive with the global human population depends, among other factors, on whether humans in large groups scale with something close to the ideal efficiency for parallelism.

In 1999, a game of chess titled "Kasparov versus The World" was played over the Internet between Garry Kasparov and a World Team in which over fifty thousand individuals participated at least once, coordinated by four young chess stars, a fifth master advising, and moves decided by majority vote with five thousand voters on a typical move. Kasparov won after four months and sixty-two moves, saying that he had never expended so much effort in his life, and later wrote a book (Kasparov and King 2000) about the game, saying, "It is the greatest game in the history of chess. The sheer number of ideas, the complexity, and the contribution it has made to chess make it the most important game ever played."

There was clearly nontrivial scaling by the contributors of the World Team—they played at a far higher skill level than their smartest individual players. But eventually Kasparov did win, and this implies that five thousand human brains (collectively representing, say, $\sim 10^{18}$ synapses) were not able to defeat Kasparov's $\sim 10^{14}$ synapses. If this seems like an unfair estimate, its unfairness may be of a type that ubiquitously characterizes human civilization's attempts to scale. Of course many of Kasparov's opponents were insufficiently skilled to be likely to make a significant contribution to suggesting or analyzing any given move; he was not facing five thousand masters. But if the World Team had possessed the probable advantages of AIs, they could have copied chess skills from one of their number to another, and thus scaled more efficiently. The fact that humans cannot do this, and that we must painstakingly and expensively reproduce the educational process for every individual who wishes to contribute to a cognitive frontier, and some our most remarkable examples cannot be duplicated by any known method of training, is one of the ways in which human populations scale less than optimally.[72]

On a more micro level, it is a truism of computer science and an important pragmatic fact of programming that processors separated by sparse communication bandwidth sometimes have trouble scaling well. When you lack the bandwidth to copy whole internal cognitive representations, computing power must be expended (wasted) to reconstruct those representations within the message receiver. It was not possible for one of Kasparov's opponents to carefully analyze an aspect of the situation and then copy and distribute that state of mind to one hundred others who could analyze slight variant

72. Update: Apparently Kasparov was reading the forums of The World during the game; in other words, he had access to their thought processes, but not the other way around. This weakens the degree of evidence substantially.

thoughts and then combine their discoveries into a single state of mind. They were limited to speech instead. In this sense it is not too surprising that $10^{14}$ synapses with high local intercommunication bandwidth and a high local skill level could defeat $10^{18}$ synapses separated by gulfs of speech and argument.

Although I expect that this section of my analysis will not be without controversy, it appears to the author to also be an important piece of data to be explained that human science and engineering seem to scale over time better than over population—an extra decade seems much more valuable than adding warm bodies.

Indeed, it appears to the author that human science scales ludicrously poorly with increased numbers of scientists, and that this is a major reason there hasn't been more relative change from 1970–2010 than from 1930–1970 despite the vastly increased number of scientists. The rate of real progress seems mostly constant with respect to time, times a small factor more or less. I admit that in trying to make this judgment I am trying to summarize an overwhelmingly distant grasp on all the fields outside my own handful. Even so, a complete halt to science or a truly exponential (or even quadratic) speedup of real progress both seem like they would be hard to miss, and the exponential increase of published papers is measurable. Real scientific progress is continuing over time, so we haven't run out of things to investigate; and yet somehow real scientific progress isn't scaling anywhere near as fast as professional scientists are being added.

The most charitable interpretation of this phenomenon would be that science problems are getting harder and fields are adding scientists at a combined pace which produces more or less constant progress. It seems plausible that, for example, Intel adds new researchers at around the pace required to keep up with its accustomed exponential growth. On the other hand, Intel actually publishes their future roadmap and is a centrally coordinated semirational agency. Scientific fields generally want as much funding as they can get from various funding sources who are reluctant to give more of it, with politics playing out to determine the growth or shrinking rate in any given year. It's hard to see how this equilibrium could be coordinated.

A moderately charitable interpretation would be that science is inherently bounded by serial causal depth and is poorly parallelizable—that the most important impacts of scientific progress come from discoveries building on discoveries, and that once the best parts of the local search field are saturated, there is little that can be done to reach destinations any faster. This is moderately uncharitable because it implies that large amounts of money are probably being wasted on scientists who have "nothing to do" when the people with the best prospects are already working on the most important problems. It is still a charitable interpretation in the sense that it implies global progress is being made around as fast as human scientists can make progress.

Both of these charitable interpretations imply that AIs expanding onto new hardware will not be able to scale much faster than human scientists trying to work in parallel, since human scientists are already working, in groups, about as efficiently as reasonably possible.

And then we have the less charitable interpretations—those which paint humanity's performance in a less flattering light.

For example, to the extent that we credit Max Planck's claim that "a new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it" (Kuhn 1962), we could expect that the process of waiting for the previous generation to die out (or rather, retire) was a serial bottleneck not affected by increased parallelism. But this would be a bottleneck of human stubbornness and aging biological brains, rather than an inherent feature of the problem space or a necessary property of rational agencies in general.

I have also wondered how it is that a ten-person startup can often appear to be around as innovative on average as a ten-thousand-person corporation. An interpretation has occurred to me which I have internally dubbed "the hero theory." This is the idea that a human organization has room for one to five "heroes" who are allowed to be important, and that other potential heroes somehow see that all hero positions are already occupied, whereupon some instinctive part of their mind informs them that there is no fame or status to be gained from heroics.[73] This theory has the advantage of explaining in a unified way why neither academic fields nor corporations seem to be able to scale "true innovation" by throwing more warm bodies at the problem, and yet are still able to scale with added time. It has the disadvantage of its mechanism not being overwhelmingly plausible. Similar phenomena might perhaps be produced by the attention span of other researchers bottlenecking through a few leaders, or by limited width of attention to funding priorities or problems. This kind of sociology is not really my field.

Diving further into the depths of cynicism, we may ask whether "science" is perhaps a process distinct from "publishing papers in journals," where our civilization understands how to reproduce the latter skill but has no systematic grasp on reproducing the former. One observes that technological progress is not (yet) dominated by China despite China graduating more PhDs than any other nation. This seems understandable if human civilization understands explicitly how to make PhDs, but the production of scientists

73. I have sometimes worried that by being "that Friendly AI guy" I have occupied the position of "Friendly AI guy" and hence young minds considering what to do with their lives will see that there is already a "Friendly AI guy" and hence not try to do this themselves. This seems to me like a very worrisome prospect, since I do not think I am sufficient to fill the entire position.

is dominated by rare lineages of implicit experts who mostly live in countries with long historical scientific traditions—and moreover, politicians or other funding agencies are bad at distinguishing the hidden keepers of the tradition and cannot selectively offer them a million dollars to move to China. In one sense this possibility doesn't say much about the true scaling factor that would apply with more scientists, but it says that a large penalty factor might apply to estimating human scaling of science by estimating scaling of publications.

In the end this type of sociology of science is not really the author's field. Nonetheless one must put probability distributions on guesses, and there is nothing especially virtuous about coming to estimates that sound respectful rather than cynical. And so the author will remark that he largely sees the data to be explained as "human science scales extremely poorly with throwing more warm bodies at a field"; and that the author generally sees the most plausible explanations as revolving around problems of the human scientific bureaucracy and process which would not necessarily hold of minds in general, especially a single AI scaling onto more hardware.

### 3.5. The Net Efficiency of Human Civilization

It might be tempting to count up 7,000,000,000 humans with 100,000,000,000 neurons and 1,000 times as many synapses firing around 100 times per second, and conclude that any rational agency wielding much fewer than $10^{26}$ computing operations per second cannot be competitive with the human species.

But to the extent that there are inefficiencies, either in individual humans or in how humans scale in groups, $10^{26}$ operations per second will not well characterize the cognitive power of the human species as a whole, as it is available to be focused on a scientific or technological problem, even relative to the characteristic efficiency of human cognitive algorithms.

A preliminary observation, that John von Neumann had a brain not much visibly larger than that of the average human, suggests that the true potential of $10^{26}$ operations per second must be bounded below by the potential of 7,000,000,000 mutually telepathic von Neumanns. Which does not seem to well characterize the power of our current civilization. Which must therefore be operating at less than perfect efficiency in the realms of science and technology.

In particular, I would suggest the following inefficiencies:

- Humans must communicate by speech and other low-bandwidth means rather than directly transferring cognitive representations, and this implies a substantial duplication of cognitive labor.

- It is possible that some professionals are systematically unproductive of important progress in their field, and the number of true effective participants must be adjusted down by some significant factor.
- Humans must spend many years in schooling before they are allowed to work on scientific problems, and this again reflects mostly duplicated cognitive labor, compared to Xeroxing another copy of Einstein.
- Human scientists do not do science twenty-four hours per day (this represents a small integer factor of reduced efficiency).
- Professional scientists do not spend all of their working hours directly addressing their scientific problems.
- Within any single human considering a scientific problem, not all of their brain can be regarded as working on that problem.
- Inefficiencies of human scientific bureaucracy may cause potentially helpful contributions to be discarded, or funnel potentially useful minds into working on problems of predictably lesser importance, etc.

One further remarks that most humans are not scientists or engineers at all, and most scientists and engineers are not focusing on the problems that an AI in the process of an intelligence explosion might be expected to focus on, like improved machine cognitive algorithms or, somewhere at the end, protein structure prediction. However, the Hansonian method of critique[74] would obviously prompt the question, "Why do you think AIs wouldn't have to spend most of their time and brainpower on subsidiary economic tasks to support themselves, just like human civilization can't afford to spend all its time on AI research?"

One reply might be that, while humans are obliged to use whole human brains to support their bodies even as they carry out relatively repetitive bits of physical or cognitive labor, an AI would be able to exploit money-earning opportunities that required straightforward cognition using a correspondingly smaller amount of computing power. The Hansonian method would then proceed to ask why there weren't many AIs bidding on such jobs and driving down the returns.[75] But in models with a localized FOOM

74. I would describe the general rule as follows: "For all supposed capabilities of AIs, ask why humans do not have the same ability. For all supposed obstacles to the human version of the ability, ask why similar obstacles would not apply to AIs." I often disagree with Hanson about whether cases of this question can be given satisfying answers, but the question itself is clearly wise and correct.

75. I would describe this rule as follows: "Check whenever someone is working on a background assumption of a localized FOOM and then consider a contrasting scenario based on many AIs of roughly

and hence one AI relatively ahead of other projects, it is very reasonable that the AI could have a much higher ratio of "computing operations doing science" to "computing operations earning money," even assuming the AI was not simply stealing its computer time. More generally, the fact that the whole human population is not mostly composed of professional scientists, working on the most important problems an AI would face in the process of going FOOM, must play a role in reducing our estimate of the net computing power required to match humanity's input into AI progress, given algorithms of roughly human-level efficiency.

All of the above factors combined may still only scratch the surface of human computational inefficiency. Our performance on integer multiplication problems is not in accordance with what a crude estimate of $10^{16}$ operations per second might lead you to expect. To put it another way, our brains do not efficiently transmit their underlying computing power to the task of integer multiplication.

Our insanely poor performance on integer multiplication clearly does not upper-bound human computational efficiency on all problems—even nonancestral problems. Garry Kasparov was able to play competitive chess against Deep Blue while Kasparov was examining two moves per second to Deep Blue's two billion moves per second, implying that Kasparov was indeed able to effectively recruit his visual cortex, temporal lobe, prefrontal cortex, cerebellum, etc., to effectively contribute large amounts of computing power in the form of parallelized pattern recognition and planning. In fact Kasparov showed amazing computational efficiency; he was able to match Deep Blue in a fashion that an *a priori* armchair reasoner probably would not have imagined possible for a mind limited to a hundred steps per second of serial depth. Nonetheless, the modern chess program Deep Rybka 3.0 is far ahead of Kasparov while running on 2.8 billion operations per second, so Kasparov's brainpower is still not being perfectly transmitted to chess-playing ability. In the end such inefficiency is what one would expect, given that Kasparov's genetic makeup was not selected over eons to play chess. We might similarly find of human scientists that, even though they are able to recruit more of their brains' power to science than to integer multiplication, they are still not using their computing operations as efficiently as a mind designed to do science—even during their moments of peak insight while they are working on that exact problem.

All these factors combined project a very different image of what an AI must do to outcompete human civilization at the task of inventing better AI algorithms or cracking protein folding than saying that the AI must compete with 7,000,000,000 humans each with $10^{11}$ neurons and $10^{14}$ synapses firing $10^2$ times per second.

---

equal ability." Here I disagree more about whether this question is really useful, since I do in fact expect a local FOOM.

By the time we are done observing that not all humans are scientists, that not all scientists are productive, that not all productive scientists are working on the problem every second, that not all professional labor is directly applicable to the cognitive problem, that cognitive labor (especially learning, or understanding ideas transmitted by speech) is often duplicated between individuals, that the fruits of nonduplicated contributions are processed by the surrounding bureaucracy with less than perfect efficiency, that humans experience significant serial bottlenecks due to their brains running on a characteristic timescale of at most $10^2$ steps per second, that humans are not telepathic, and finally that the actual cognitive labor applied to the core cognitive parts of scientific problems during moments of peak insight will be taking place at a level of inefficiency somewhere between "Kasparov losing at chess against Deep Rybka's 2.8 billion operations/second" and "Kasparov losing at integer multiplication to a pocket calculator" . . .

. . . the effective computing power of human civilization applied to the relevant problems may well be within easy range of what a moderately well-funded project could simply buy for its AI, without the AI itself needing to visibly earn further funding.

Frankly, my suspicion is that by the time you're adding up *all* the human inefficiencies, then even without much in the way of fundamentally new and better algorithms—just boiling down the actual cognitive steps required by the algorithms we already use—well, it's actually quite low, I suspect.[76]

And this probably has a substantial amount to do with why, in practice, I think a moderately well-designed AI could overshadow the power of human civilization. It's not just about abstract expectations of future growth, it's a sense that the net cognitive ability of human civilization is not all that impressive once all the inefficiencies are factored in. Someone who thought that $10^{26}$ operations per second was actually a good proxy measure of the magnificent power of human civilization might think differently.

### 3.6. Returns on Cumulative Evolutionary Selection Pressure

I earlier claimed that we have seen no signs of diminishing cognitive returns to cumulative natural selection. That is, it didn't take one-tenth as long to go from *Australopithecus* to *Homo erectus* as it did from *Homo erectus* to *Homo sapiens*. The alert reader may protest, "Of course the *erectus–sapiens* interval isn't ten times as long as the *Australopithecus–erectus*

76. Though not as low as if all the verbal thoughts of human scientists could be translated into first-order logic and recited as theorems by a ridiculously simple AI engine, as was briefly believed during the early days. If the claims made by the makers of BACON (Langley, Bradshaw, and Zytkow 1987) or the Structure Mapping Engine (Falkenhainer and Forbus 1990) were accurate models of human cognitive reasoning, then the Scientific Revolution up to 1900 would have required on the order of perhaps $10^6$ cognitive operations *total*. We agree however with Chalmers, French, and Hofstadter (1992) that this is not a good model. So not quite *that* low.

interval, you just picked three named markers on the fossil record that didn't happen to have those relative intervals." Or, more charitably: "Okay, you've shown me some named fossils A, B, C with 3.2 million years from A to B and then 1.8 million years from B to C. What you're really claiming is that there wasn't ten times as much cognitive improvement from A to B as from B to C. How do you know that?"

To this I could reply by waving my hands in the direction of the details of neuroanthropology,[77] and claiming that the observables for throat shapes (for language use), preserved tools and campfires, and so on, just sort of *look* linear—or moderately superlinear, but at any rate not sublinear. A graph of brain sizes with respect to time may be found here (Calvin 2004, chap. 5). And despite the inferential distance from "brain size" to "increasing marginal fitness returns on brain size" to "brain algorithmic improvements"—nonetheless, the chart looks either linear or moderately superlinear.

More broadly, another way of framing this is to ask what the world should look like if there *were* strongly decelerating returns to evolutionary optimization of hominids.[78]

I would reply that, first of all, it would be very surprising to see a world whose cognitive niche was dominated by just one intelligent species. Given sublinear returns on cumulative selection for cognitive abilities, there should be other species that mostly catch up to the leader. Say, evolving sophisticated combinatorial syntax from protolanguage should have been a much more evolutionarily expensive proposition than just producing protolanguage, due to the decelerating returns.[79] And then, in the long time it took hominids to evolve complex syntax from protolanguage, chimpanzees should have caught up and started using protolanguage. Of course, evolution does not always recapitulate the same outcomes, even in highly similar species. But in general, sublinear cognitive returns to evolution imply that it would be surprising to see one species get far ahead of all others; there should be nearly even competitors in the process of catching

77. Terrence Deacon's (1997) *The Symbolic Species* is notionally about a theory of human general intelligence which I believe to be quite mistaken, but the same book is incidentally an excellent popular overview of cognitive improvements over the course of hominid evolution, especially as they relate to language and abstract reasoning.

78. At the Center for Applied Rationality, one way of training empiricism is via the Monday-Tuesday game. For example, you claim to believe that cellphones work via "radio waves" rather than "magic." Suppose that on Monday cellphones worked via radio waves and on Tuesday they worked by magic. What would you be able to *see* or *test* that was different between Monday and Tuesday?

Similarly, here we are asking, "On Monday there are linear or superlinear returns on cumulative selection for better cognitive algorithms. On Tuesday the returns are strongly sublinear. How does the world look different on Monday and Tuesday?"

To put it another way: If you have strongly concluded X, you should be able to easily describe how the world would look very different if not-X, or else how did you conclude X in the first place?

79. For an explanation of "protolanguage" see Bickerton (2009).

up. (For example, we see millions of species that are poisonous, and no one species that has taken over the entire "poison niche" by having far better poisons than its nearest competitor.)

But what if there were hugely increased *selection pressures* on intelligence within hominid evolution, compared to chimpanzee evolution? What if, over the last 1.8 million years since *Homo erectus*, there was a thousand times as much selection pressure on brains in particular, so that the cumulative optimization required to go from *Homo erectus* to *Homo sapiens* was in fact comparable with all the evolution of brains since the start of multicellular life?

There are mathematical limits on total selection pressures within a species. However, rather than total selection pressure increasing, it's quite plausible for selection pressures to suddenly focus on one characteristic rather than another. Furthermore, this has almost certainly been the case in hominid evolution. Compared to, say, scorpions, a competition between humans is much more likely to revolve around who has the better brain than around who has better armor plating. More variance in a characteristic which covaries with fitness automatically implies increased selective pressure on that characteristic.[80] Intuitively speaking, the more interesting things hominids did with their brains, the more of their competition would have been about cognition rather than something else.

And yet human brains actually do seem to look a lot like scaled-up chimpanzee brains—there's a larger prefrontal cortex and no doubt any number of neural tweaks, but the gross brain anatomy has changed hardly at all.

In terms of pure *a priori* evolutionary theory—the sort we might invent if we were armchair theorizing and had never seen an intelligent species evolve—it wouldn't be too surprising to imagine that a planet-conquering organism had developed a new complex brain from scratch, far more complex than its nearest competitors, after that organ suddenly became the focus of intense selection sustained for millions of years.

But in point of fact we don't see this. Human brains look like scaled-up chimpanzee brains, rather than mostly novel organs.

Why is that, given the persuasive-sounding prior argument for how there could have plausibly been thousands of times more selection pressure per generation on brains, compared to previous eons?

Evolution is strongly limited by serial depth, even though many positive mutations can be selected on in parallel. If you have an allele B which is only advantageous in the presence of an allele A, it is necessary that A rise to universality, or at least prevalence, within the gene pool before there will be significant selection pressure favoring B. If

80. For a mathematical quantification see Price's Equation.

C depends on both A and B, both A and B must be highly prevalent before there is significant pressure favoring C.[81] Within a sexually reproducing species where any genetic variance is repeatedly scrambled, complex machines will be mostly composed of a deep, still pool of complexity, with a surface froth of non-interdependent improvements being selected on at any given point. Intensified selection pressures may increase the speed at which individually positive alleles rise to universality in the gene pool, or allow for selecting on more non-interdependent variations in parallel. But there's still an important sense in which the evolution of complex machinery is strongly limited by serial depth.

So even though it is extremely plausible that hominids experienced greatly intensified selection on brains versus other organismal characteristics, it still isn't surprising that human brains look mostly like chimpanzee brains when there have only been a few hundred thousand generations separating us.

Nonetheless, the moderately superlinear increase in hominid brain sizes over time could easily accommodate strictly linear returns on cumulative selection pressures, with the seeming acceleration over time being due only to increased selection pressures on intelligence. It would be surprising for the cognitive "returns on cumulative selection pressure" *not* to be beneath the curve for "returns on cumulative time."

I was recently shocked to hear about claims for molecular evidence that rates of genetic change may have increased *one hundred-fold* among humans since the start of agriculture (Hawks et al. 2007). Much of this may have been about lactose tolerance, melanin in different latitudes, digesting wheat, etc., rather than positive selection on new intelligence-linked alleles. This still allows some potential room to attribute some of humanity's gains over the last ten thousand years to literal evolution, not just the accumulation of civilizational knowledge.

But even a literally hundredfold increase in rates of genetic change does not permit cognitive returns per individual mutation to have fallen off significantly over the course of hominid evolution. The mathematics of evolutionary biology says that a single mutation event which conveys a fitness advantage of $s$, in the sense that the average fitness of its bearer is $1 + s$ compared to a population average fitness of 1, has a $2s$ probability of spreading through a population to fixation; and the expected fixation time is $2 \ln(N)/s$ generations, where $N$ is total population size. So if the fitness advantage per positive mutation falls low enough, not only will that mutation take a very large number of

81. Then along comes A* which depends on B and C, and now we have a complex interdependent machine which fails if you remove any of A*, B, or C. Natural selection naturally and automatically produces "irreducibly" complex machinery along a gradual, blind, locally hill-climbing pathway.

generations to spread through the population, it's very likely not to spread at all (even if the mutation independently recurs many times).

The possibility of increased selection pressures should mainly lead us to suspect that there are huge cognitive gaps between humans and chimpanzees which resulted from merely linear returns on cumulative optimization—there was a lot more optimization going on, rather than small amounts of optimization yielding huge returns. But we can't have a small cognitive gap between chimps and humans, a large amount of cumulative selection, and fitness returns on individual mutations strongly diminishing, because in this scenario we wouldn't get much evolution, period. The possibility of increased rates of genetic change does not actually imply room for cognitive algorithms becoming "harder to design" or "harder to improve upon" as the base level grows more sophisticated. Returns on single positive mutations are lower-bounded by the logic of natural selection.

If you think future molecular genetics might reveal these sorts of huge selection pressures in the historical record, you should consistently think it plausible (though perhaps not certain) that humans are vastly smarter than chimps (contrary to some arguments in the opposite direction, considered in section 3.2). There is room for the mind-design distance from *Homo erectus* to *Homo sapiens* to be significant compared to, say, the mind-design distance from mouse to *Australopithecus*, contrary to what the relative time intervals in the fossil record would suggest.

To wedge diminishing cognitive returns on evolution into this model—without contradicting basic evolutionary points about how sufficiently small fitness advantages take huge amounts of time to fixate, or more likely don't fixate at all—we would have to suppose that small cognitive advantages were somehow providing outsize fitness advantages (in a way irrelevant to returns on cognitive reinvestment for AIs trying to improve themselves). To some degree, "inflated fitness advantages" occur in theories of runaway sexual selection (where everyone tries to mate with whoever seems even nominally smartest). To whatever extent such sexual selection was occurring, we should decrease our estimate of the sort of cognitively produced fitness advantage that would carry over to a machine intelligence trying to work on the protein folding problem (where you do not get an outsized prize for being only slightly better).

I would nonetheless say that, at the end of the day, it takes a baroque interpretation of the graph of brain sizes with respect to time, to say nothing of the observed cognitive gap between humans and chimps, before you can get *diminishing* returns on cumulative natural selection out of observed bioanthropology. There's some room for short recent time intervals to expand into large amounts of cumulative selection pressure, but this mostly means that we don't need to postulate increasing returns on each positive mu-

tation to account for apparently superlinear historical progress.[82] On the whole, there is not much room to postulate that evolutionary history is telling us about decreasing cognitive returns to cumulative natural selection.

### 3.7. Relating Curves of Evolutionary Difficulty and Engineering Difficulty

What if creating human intelligence was easy for natural selection but will be hard for human engineers?

The power of natural selection is often romanticized—for example, because of cultural counterpressures in the United States to religions that try to falsely downplay the power of natural selection. Even some early biologists made such errors, although mostly before George C. Williams (1966) and the revolution of the 1960s, which spawned a very clear, often mathematically precise, picture of the capabilities and characteristic design processes of natural selection. Today we can in many respects quantify with simple equations the statement that natural selection is slow, stupid, and blind: a positive mutation of fitness $1 + s$ will require $2\ln(\text{population})/s$ generations to fixate and has only a $2s$ probability of doing so at all.[83]

Evolution has invented the freely rotating wheel on only a tiny handful of occasions in observed biology. Freely rotating wheels are in fact highly efficient—that is why they appear in ATP synthase, a molecule which may have been selected more heavily for near-perfect efficiency than almost anything else in biology. But (especially once we go from self-assembling molecules to organs which must be grown from tissue) it's hard to come by intermediate evolutionary forms along the way to a freely rotating wheel. Evolution cannot develop intermediate forms *aiming* for a freely rotating wheel, and it almost never locally hill-climbs into that design. This is one example of how human engineers, who can hold whole designs in their imagination and adjust them in response to imagined problems, can easily access areas of design space which evolution almost never enters.

We should strongly expect that point mutation, random recombination, and statistical selection would hit bottlenecks in parts of the growth curve where deliberate foresight, consequentialist back-chaining, and learned abstraction would carry steadily onward—rather than the other way around. Difficulty curves for intelligent engineers

82. To be clear, increasing returns per positive mutation would imply that improving cognitive algorithms became easier as the base design grew more sophisticated, which would imply accelerating returns to constant optimization. This would be one possible explanation for the seemingly large gains from chimps to humans, but the fact that selection pressures almost certainly increased, and may have increased by quite a lot, means we cannot strongly conclude this.

83. Imagine if each 2% improvement to car engines, since the time of the Model T, had required a thousand generations to be adopted and had only a 4% chance of being adopted at all.

should be bounded upward by the difficulty curves for the processes of natural selection (where higher difficulty represents lower returns on cumulative investment). Evolution does have a significant head start. But while trying to catch up with millions of years of cumulative evolutionary optimization sounds intimidating at first, it becomes less intimidating once you calculate that it takes 875 generations for a gene conveying a 3% fitness advantage to spread through a population of five hundred thousand individuals.

We can't expect the difficulty curves for intelligent engineering and natural selection to be the same. But we can reasonably relate them by saying that the difficulty curve for intelligent engineering should stay below the corresponding curve for natural selection, but that natural selection has a significant head start on traversing this curve.

Suppose we accept this relation. Perhaps we still can't conclude very much in practice about AI development times. Let us postulate that it takes eighty years for human engineers to get AI at the level of *Homo erectus*. Plausibly *erectus*-level intelligence is still not smart enough for the AI to contribute significantly to its own development (though see section 3.10).[84] Then, if it took eighty years to get AI to the level of *Homo erectus*, would it be astonishing for it to take another ninety years of engineering to get to the level of *Homo sapiens?*

I would reply, "Yes, I would be astonished, because even after taking into account the possibility of recently increased selection pressures, it still took far more evolutionary time to get to *Homo erectus* from scratch than it took to get from *Homo erectus* to *Homo sapiens*." If natural selection didn't experience a sharp upward difficulty gradient after reaching the point of *Homo erectus*, it would be astonishing to find that human engineering could reach *Homo erectus*-level AIs (overcoming the multi-hundred-million-year cumulative lead natural selection had up until that point) but that human engineering then required *more* effort to get from there to a *Homo sapiens* equivalent.

But wait: the human-engineering growth curve could be bounded below by the evolutionary curve while still having a different overall shape. For instance it could be that all the steps up to *Homo erectus* are much easier for human engineers than evolution—that the human difficulty curve over this region is far below the evolutionary curve—and then the steps from *Homo erectus* to *Homo sapiens* are only slightly easier for human engineers. That is, the human difficulty curve over this region is moderately below the evolutionary curve. Or to put it another way, we can imagine that *Homo erectus*

84. The reason this statement is not obvious is that an AI with *general* intelligence roughly at the level of *Homo erectus* might still have outsized abilities in computer programming—much as modern AIs have poor cross-domain intelligence, and yet there are still specialized chess AIs. Considering that blind evolution was able to build humans, it is not obvious that a sped-up *Homo erectus* AI with specialized programming abilities could not improve itself up to the level of *Homo sapiens*.

was "hard" for natural selection and getting from there to *Homo sapiens* was "easy," while both processes will be "easy" for human engineers, so that both steps will take place in eighty years each. Thus, the statement "Creating intelligence will be much easier for human engineers than for evolution" could imaginably be true in a world where "It takes eighty years to get to *Homo erectus* AI and then another ninety years to get to *Homo sapiens* AI" is also true.

But one must distinguish possibility from probability. In probabilistic terms, I would be astonished if that actually happened, because there we have no observational reason to suppose that the relative difficulty curves actually look like that; specific complex irregularities with no observational support have low prior probability. When I imagine it concretely I'm also astonished: If you can build *Homo erectus* you can build the cerebral cortex, cerebellar cortex, the limbic system, the temporal lobes that perform object recognition, and so on. Human beings and chimpanzees have the vast majority of their neural architectures in common—such features have not diverged since the last common ancestor of humans and chimps. We have some degree of direct observational evidence that human intelligence is the icing on top of the cake that is chimpanzee intelligence. It would be surprising to be able to build that much cake and then find ourselves unable to make a relatively small amount of icing. The 80–90 hypothesis also requires that natural selection would have had an easier time building more sophisticated intelligences—equivalently, a harder time building less sophisticated intelligences—for reasons that wouldn't generalize over to human engineers, which further adds to the specific unsupported complex irregularity.[85]

In general, I think we have specific reason to suspect that difficulty curves for natural selection bound above the difficulty curves for human engineers, and that humans will be able to access regions of design space blocked off from natural selection. I would expect early AIs to be in some sense intermediate between humans and natural selection in this sense, and for sufficiently advanced AIs to be further than humans along the same spectrum. Speculations which require specific unsupported irregularities of the relations

85. By the method of imaginary updates, suppose you told me, "Sorry, I'm from the future, and it so happens that it really *did* take X years to get to the *Homo erectus* level and then another X years to get to the *Homo sapiens* level." When I was done being shocked, I would say, "Huh. I guess there must have been some way to get the *equivalent* of *Homo erectus* performance without building anything remotely like an actual *Homo erectus*, in a way that didn't generalize over to doing things *Homo sapiens* can do." (We already have AIs that can surpass human performance at chess, but in a way that's not at all like the way humans solve the problem and that doesn't generalize to other human abilities. I would suppose that *Homo erectus*-level performance on most problems had been similarly obtained.) It would still be just too surprising for me to believe that you could literally build a *Homo erectus* and then have that much trouble getting to *Homo sapiens*.

between these curves should be treated as improbable; on the other hand, outcomes which would be yielded by many possible irregularities are much more probable, since the relations are bound to be irregular somewhere. It's possible that further analysis of this domain could yield more specific statements about expected relations between human engineering difficulty and evolutionary difficulty which would be relevant to AI timelines and growth curves.

### 3.8. Anthropic Bias in Our Observation of Evolved Hominids

The observation "intelligence evolved" may be misleading for anthropic reasons: perhaps evolving intelligence is incredibly difficult, but on all the planets where it doesn't evolve, there is nobody around to observe its absence.

Shulman and Bostrom (2012) analyzed this question and its several possible answers given the present state of controversy regarding how to reason about anthropic probabilities. Stripping out a number of caveats and simplifying, it turns out that—under assumptions that yield any adjustment at all for anthropic bias—the main conclusion we can draw is a variant of Hanson's (1998c) conclusion: if there are several "hard steps" in the evolution of intelligence, then planets on which intelligent life does evolve should expect to see the hard steps spaced about equally across their history, regardless of each step's relative difficulty.

Suppose a large population of lockpickers are trying to solve a series of five locks in five hours, but each lock has an average solution time longer than five hours—requiring ten hours or a hundred hours in the average case. Then the few lockpickers lucky enough to solve every lock will probably see the five locks distributed randomly across the record. Conditioning on the fact that a lockpicker was lucky enough to solve the five locks at all, a hard lock with an average solution time of ten hours and a hard lock with an average solution time of one hundred hours will have the same expected solution times selecting on the cases where all locks were solved.[86]

This in turn means that "self-replicating life comes into existence" or "multicellular organisms arise" are plausible hard steps in the evolution of intelligent life on Earth, but the time interval from *Australopithecus* to *Homo sapiens* is too short to be a plausible hard step. There might be a hard step along the way to first reaching *Australopithecus* intelligence,

86. I think a legitimate simplified illustration of this result is that, given a solution time for lock A evenly distributed between 0 hours and 200 hours and lock B with a solution time evenly distributed between 0 hours and 20 hours, then *conditioning* on the fact that A and B were both successfully solved in a total of 2 hours, we get equal numbers for "the joint probability that A was solved in 1.5–1.6 hours and B was solved in 0.4–0.5 hours" and "the joint probability that A was solved in 0.4–0.5 hours and B was solved in 1.5–1.6 hours," even though in both cases the probability for A being solved that fast is one-tenth the probability for B being solved that fast.

but from chimpanzee-equivalent intelligence to humans was apparently smooth sailing for natural selection (or at least the sailing was probably around as smooth or as choppy as the "naive" perspective would have indicated before anthropic adjustments). Nearly the same statement could be made about the interval from mouse-equivalent ancestors to humans, since fifty million years is short enough for a hard step to be improbable, though not quite impossible. On the other hand, the gap from spiders to lizards might more plausibly contain a hard step whose difficulty is hidden from us by anthropic bias.

What does this say about models of the intelligence explosion?

Difficulty curves for evolution and for human engineering cannot reasonably be expected to move in lockstep. Hard steps for evolution are not necessarily hard steps for human engineers (recall the case of freely rotating wheels). Even if there has been an evolutionarily hard step on the road to mice—a hard step that reduced the number of planets with mice by a factor of $10^{50}$, emptied most galactic superclusters of mice, and explains the Great Silence we observe in the night sky—it might still be something that a human engineer can do without difficulty.[87] If natural selection requires $10^{100}$ tries to do something but eventually succeeds, the problem still can't be that hard in an absolute sense, because evolution is still pretty stupid.

There is also the possibility that we could reverse-engineer actual mice. I think the role of reverse-engineering biology is often overstated in Artificial Intelligence, but if the problem turns out to be incredibly hard for mysterious reasons, we do have mice on hand.

Thus an evolutionarily hard step would be relatively unlikely to represent a *permanent* barrier to human engineers.

All this only speaks of a barrier along the pathway to producing mice. One reason I don't much modify my model of the intelligence explosion to compensate for possible anthropic bias is that a humanly difficult barrier below the mouse level looks from the outside like, "Gosh, we've had lizard-equivalent AI for twenty years now and we still can't get to mice, we may have to reverse-engineer actual mice instead of figuring this out on our own."[88] But the advice from anthropics is that the road from mice to humans

87. It's interesting to note that human engineers have not yet built fully self-replicating systems, and the initial emergence of self-replication is a plausible hard step. On the other hand, the emergence of complex cells (eukaryotes) and then multicellular life are both plausible hard steps requiring about a billion years of evolution apiece, and human engineers don't seem to have run into any comparable difficulties in making complex things with complex parts.

88. It's hard to eyeball this sort of thing, but I don't see any particular signs that AI has gotten stuck at any particular point so far along the road to mice. To observers outside the field, AI may appear bottlenecked because in normal human experience, the scale of intelligence runs from "village idiot" to "Einstein," and so it intuitively appears that AI is stuck and unmoving below the "village idiot level." If

is no more difficult than it looks, so a "hard step" which slowed down an intelligence explosion in progress would presumably have to strike before that intelligence explosion hit the mouse level.[89] Suppose an intelligence explosion could in fact get started beneath the mouse level—perhaps a specialized programming AI with sub-mouse general intelligence and high serial speeds might be able make significant self-improvements. Then from the outside we would see something like, "Huh, we can build these relatively dumb specialized AIs that seem to get significant mileage out of recursive self-improvement, but then everything we build bottlenecks around the same sub-mouse level."

If we tried hard to derive policy advice from this anthropic point, it might say: "If tomorrow's AI researchers can build relatively dumb self-modifying systems that often manage to undergo long chains of significant self-improvement with reinvested returns, and they all get stuck at around the same point somewhere below mouse-level general intelligence, then it's possible that this point is the "hard step" from evolutionary history, rather than a place where the difficulty curve permanently slopes upward. You should potentially worry about the first AI that gets pushed past this big sticking point, because once you do get to mice, it may be an easy journey onward from there." I'm not sure I'd have very much confidence in that advice—it seems to have been obtained via a complicated argument and I don't see a good way to simplify the core idea. But since I wouldn't otherwise expect this kind of bottlenecking to be uniform across many different AI systems, that part is arguably a unique prediction of the hard-step model where some small overlooked lock actually contains a thousand cosmic hours of average required solution time.

For the most part, though, it appears to me that anthropic arguments do not offer very detailed advice about the intelligence explosion (and this is mostly to be expected).

### 3.9. Local versus Distributed Intelligence Explosions

A key component of the debate between Robin Hanson and myself was the question of locality. Consider: If there are increasing returns on knowledge given constant human

---

you are properly appreciating a scale that runs from "rock" at zero to "bacterium" to "spider" to "lizard" to "mouse" to "chimp" to "human," then AI seems to be moving along at a slow but steady pace. (At least it's slow and steady on a human R&D scale. On an evolutionary scale of time, progress in AI has been unthinkably, blindingly fast over the past sixty-year instant.) The "hard step" theory does say that we might expect some further mysterious bottleneck, short of mice, to a greater degree than we would expect if not for the Great Silence. But such a bottleneck might still not correspond to a huge amount of time for human engineers.

89. A further complicated possible exception is if we can get far ahead of lizards in some respects, but are missing one vital thing that mice do. Say, we already have algorithms which can find large prime numbers much faster than lizards, but still can't eat cheese.

brains—this being the main assumption that many non-intelligence-explosion, general technological hypergrowth models rely on, with said assumption seemingly well-supported by exponential[90] technology-driven productivity growth[91]—then why isn't the leading human nation vastly ahead of the runner-up economy? Shouldn't the economy with the most knowledge be rising further and further ahead of its next-leading competitor, as its increasing returns compound?

The obvious answer is that knowledge is not contained within the borders of one country: improvements within one country soon make their way across borders. China is experiencing greater growth per annum than Australia, on the order of 8% versus 3% RGDP growth.[92] This is not because technology development in general has diminishing marginal returns. It is because China is experiencing very fast knowledge-driven growth as it catches up to already-produced knowledge that it can cheaply import.

Conversely, hominids moved further and further ahead of chimpanzees, who fell further behind rather than catching up, because hominid genetic innovations did not make it into the chimpanzee gene pool. We can speculate about how brain improvements might have led to increased cognitive returns on further improvements, or how cognitive improvements might have increased selection pressures surrounding intelligence, creating a positive feedback effect in hominid evolution. But this still would not have

90. The word "exponential" does not mean "fast"; it means a solution of the differential equation $y' = ky$. The "Great Stagnation" thesis revolves around the claim that total-factor productivity growth in developed countries was running at around 0.75% per annum during the twentieth century until it dropped to 0.25% per annum in the mid-1970s (Cowen 2011). This is not *fast*, but it is exponential.

91. I suspect that uncertainty about how fast humans can compound technological progress is not the question that dominates uncertainty about growth rates in the intelligence explosion, so I don't talk much about the curve of human technological progress one way or another, except to note that there is some. For models of technological hypergrowth that only try to deal in constant human brains, such details are obviously of much greater interest.

Personally I am agnostic, leaning skeptical, about technological hypergrowth models that don't rely on cognitive reinvestment. I suspect that if you somehow had constant human brains—no genetic engineering of humans, no sixty-four-node clustered humans using brain-computer interfaces, no faster researchers, no outsized cognitive returns from superintelligent AI, no molecular nanotechnology, and nothing else that permitted cognitive reinvestment—then the resulting scenario might actually look pretty normal for a century; it is plausible to me that there would be roughly the same amount of technology-driven change from 2000–2100 as from 1900–2000. (I would be open to hearing why this is preposterous.)

92. Japan is possibly the country with the most advanced technology per capita, but their economic growth has probably been hampered by Japanese monetary policy. Scott Sumner likes Australia's monetary policy, so I'm comparing China to Australia for purposes of comparing growth rates in developing vs. developed countries.

caused hominids to pull far ahead of other primates, if hominid improvements had been spreading to primates via horizontal gene transmission.[93]

Thus we can sketch two widely different possible scenarios for an intelligence explosion, at opposite extremes along multiple dimensions, as follows:[94]

**Extremely local takeoff:**

- Much like today, the diversity of advanced AI architectures is so great that there is very little trading of cognitive content between projects. It's easier to download a large dataset, and have your AI relearn the lessons of that dataset within its own cognitive representation, than to trade cognitive content between different AIs. To the extent that AIs other than the most advanced project can generate self-improvements at all, they generate modifications of idiosyncratic code that can't be cheaply shared with any other AIs.
- The leading projects do not publish all or even most of their research—whether for the same reasons hedge funds keep their sauces secret, or for the same reason Leo Szilard didn't immediately tell the world about fission chain reactions.
- There is a relatively small number of leading projects.
- The first AI to touch the intelligence explosion reaches $k > 1$ due to a basic algorithmic improvement that hasn't been shared with any other projects.
- The AI has a sufficiently clean architecture that it can scale onto increasing amounts of hardware while remaining as a unified optimization process capable of pursuing coherent overall goals.
- The AI's self-improvement, and eventual transition to rapid infrastructure, involves a large spike in capacity toward the latter end of the curve (as superintelligence is achieved, or as protein structure prediction is cracked sufficiently to build later stages of nanotechnology). This vastly amplifies the AI's cognitive and technological lead time over its nearest competitor. If the nearest competitor was previously only seven days behind, these seven days have now been amplified into a technological gulf enabling the leading AI to shut down, sandbox, or restrict the growth

93. Theoretically, genes can sometimes jump this sort of gap via viruses that infect one species, pick up some genes, and then infect a member of another species. Speaking quantitatively and practically, the amount of gene transfer between hominids and chimps was approximately zero so far as anyone knows.

94. Again, neither of these possibilities should be labeled "good" or "bad"; we should make the best of whatever reality we turn out to live in, whatever the settings of the hidden variables.

of any competitors it wishes to fetter. The final result is a Bostrom (2006)-style "singleton."

**Extremely global takeoff:**

- The emergence of good, successful machine intelligence techniques greatly winnows the plethora of visionary prototypes we see nowadays (Hanson 2008b). AIs are similar enough that they can freely trade cognitive content, code tweaks, and algorithmic improvements.
- There are many, many such AI projects.
- The vast majority of "improvement" pressure on any single machine intelligence derives from the total global economy of machine intelligences or from academic AI researchers publishing their results, not from that AI's internal self-modifications. Although the global economy of machine intelligences is getting high returns on cognitive investments, no single part of that economy can go FOOM by itself.
- Any sufficiently large machine intelligence is forced by lack of internal bandwidth to split into pieces, which then have their own local goals and do not act as a well-coordinated whole.
- The benefit that an AI can derive from local use of an innovation is very small compared to the benefit that it can get from selling the innovation to many different AIs. Thus, very few innovations are kept secret. (The same reason that when Stephen King writes a novel, he sells the novel to hundreds of thousands of readers and uses the proceeds to buy more books, instead of just keeping the novel to himself.)
- Returns on investment for machine intelligences which fall behind automatically increase as the machine is enabled to "catch up" on cheaper knowledge (much as China is growing faster than Australia). Also, leading agencies do not eliminate laggards or agglomerate them (the way strong countries used to conquer weak countries).
- Nobody knows how to 90%-solve the protein structure prediction problem before somebody else knows how to 88%-solve the protein structure prediction problem; relative leads are small. Even technologies like molecular nanotech appear gradually and over many different places at once, with much sharing/selling of innovations and laggards catching up; relative leads are not significantly amplified by the transition.
- The end result has a lot of trade and no global coordination. (This is not necessarily a good thing. See Hanson's [2008d] rapacious hardscrapple frontier folk.)

These two extremes differ along many dimensions that could potentially fail to be correlated. Note especially that *sufficiently* huge returns on cognitive reinvestment will produce winner-take-all models and a local FOOM regardless of other variables. To make this so extreme that even I don't think it's plausible, if there's a simple trick that lets you get molecular nanotechnology and superintelligence five seconds after you find it,[95] then it's implausible that the next runner-up will happen to find it in the same five-second window.[96] Considering five seconds as a literal time period rather than as a metaphor, it seems clear that sufficiently high returns on reinvestment produce singletons almost regardless of other variables. (Except possibly for the stance "sufficiently large minds must inevitably split into bickering components," which could hold even in this case.[97])

It should also be noted that the "global" scenario need not include all of the previous civilization inside its globe. Specifically, biological humans running on 200 Hz neurons with no read-write ports would tend to be left out of the FOOM, unless some AIs are specifically motivated to help humans as a matter of final preferences. Newly discovered cognitive algorithms do not easily transfer over to human brains with no USB ports. Under this scenario humans would be the equivalent of emerging countries with dreadfully restrictive laws preventing capital inflows, which can stay poor indefinitely. Even if it were possible to make cognitive improvements cross the "human barrier," it seems unlikely to offer the highest natural return on investment compared to investing in a fellow machine intelligence. In principle you can evade the guards and sneak past the borders of North Korea and set up a convenience store where North Koreans can buy the same goods available elsewhere. But this won't be the *best* way to invest your money—not unless you care about North Koreans as a matter of final preferences over terminal outcomes.[98]

---

95. À la *The Metamorphosis of Prime Intellect* by Roger Williams (2002).

96. A rational agency has no convergent instrumental motive to sell a *sufficiently powerful, rapidly reinvestable* discovery to another agency of differing goals, because even if that other agency would pay a billion dollars for the discovery in one second, you can get a larger fraction of the universe to yourself and hence even higher total returns by keeping mum for the five seconds required to fully exploit the discovery yourself and take over the universe.

97. This stance delves into AI-motivational issues beyond the scope of this paper. I will quickly note that the Orthogonality Thesis opposes the assertion that any "mind" must develop indexically selfish preferences which would prevent coordination, even if it were to be granted that a "mind" has a maximum individual size. Mostly I would tend to regard the idea as anthropomorphic—humans have indexically selfish preferences and group conflicts for clear evolutionary reasons, but insect colonies with unified genetic destinies and whole human brains (likewise with a single genome controlling all neurons) don't seem to have analogous coordination problems.

98. Our work on decision theory also suggests that the best coordination solutions for computer-based minds would involve knowledge of each others' source code or crisp adoption of particular crisp decision

The highly local scenario obviously offers its own challenges as well. In this case we mainly want the lead project at any given point to be putting sufficiently great efforts into "Friendly AI."[99] In the highly global scenario we get incremental improvements by having only some AIs be human-Friendly,[100] while the local scenario is winner-take-all. (But to have one AI of many be Friendly does still require that someone, somewhere solve the associated technical problem before the global AI ecology goes FOOM; and relatively larger returns on cognitive reinvestment would narrow the amount of time available to do solve that problem.)

My own expectations lean toward scenario (1)—for instance, I usually use the singular rather than plural when talking about that-which-goes-FOOM. This is mostly because I expect large enough returns on cognitive reinvestment to dominate much of my uncertainty about other variables. To a lesser degree I am impressed by the diversity and incompatibility of modern approaches to machine intelligence, but on this score I respect Hanson's argument for why this might be expected to change. The rise of open-source chess-playing programs has undeniably led to faster progress due to more sharing of algorithmic improvements, and this combined with Hanson's argument has shifted me significantly toward thinking that the ecological scenario is not completely unthinkable.

It's also possible that the difference between local-trending and global-trending outcomes is narrow enough to depend on policy decisions. That is, the settings on the hidden variables might turn out to be such that, if we wanted to see a "Friendly singleton" rather than a Hansonian "rapacious hardscrapple frontier" of competing AIs, it would be feasible to create a "nice" project with enough of a research advantage (funding, computing resources, smart researchers) over the next runner-up among non-"nice"

---

theories. Here it is much harder to verify that a human is trustworthy and will abide by their agreements, meaning that humans might "naturally" tend to be left out of whatever coordination equilibria develop among machine-based minds, again unless there are specific final preferences to include humans.

99. The Fragility of Value subthesis of Complexity of Value implies that solving the Friendliness problem is a mostly satisficing problem with a sharp threshold, just as dialing nine-tenths of my phone number correctly does not connect you to someone 90% similar to Eliezer Yudkowsky. If the fragility thesis is correct, we are not strongly motivated to have the lead project be 1% better at Friendly AI than the runner-up project; rather we are strongly motivated to have it do "well enough" (though this should preferably include some error margin). Unfortunately, the Complexity of Value thesis implies that "good enough" Friendliness involves great (though finite) difficulty.

100. Say, one Friendly AI out of a million cooperating machine intelligences implies that one millionth of the universe will be used for purposes that humans find valuable. This is actually quite a lot of matter and energy, and anyone who felt diminishing returns on population or lifespan would probably regard this scenario as carrying with it most of the utility.

competitors to later become a singleton.[101] This could be true even in a world where a global scenario would be the default outcome (e.g., from open-source AI projects) so long as the hidden variables are not too heavily skewed in that direction.

### 3.10. Minimal Conditions to Spark an Intelligence Explosion

I. J. Good spoke of the intelligence explosion beginning from an "ultraintelligence . . . a machine that can far surpass all the intellectual activities of any man however clever." This condition seems sufficient, but far more than necessary.

Natural selection does not far surpass every intellectual capacity of any human—it cannot write learned papers on computer science and cognitive algorithms—and yet it burped out a human-equivalent intelligence anyway.[102] Indeed, natural selection built humans via an optimization process of point mutation, random recombination, and statistical selection—without foresight, explicit world-modeling, or cognitive abstraction. This quite strongly upper-bounds the algorithmic sophistication required, in principle, to output a design for a human-level intelligence.

Natural selection did use vast amounts of computational brute force to build humans. The "naive" estimate is that natural selection searched in the range of $10^{30}$ to $10^{40}$ organisms before stumbling upon humans (Baum 2004). Anthropic considerations (did other planets have life but not intelligent life?) mean the real figure might be almost arbitrarily higher (see section 3.8).

There is a significant subfield of machine learning that deploys evolutionary computation (optimization algorithms inspired by mutation/recombination/selection) to try to solve real-world problems. The toolbox in this field includes "improved" genetic algorithms which, at least in some cases, seem to evolve solutions orders of magnitude faster than the first kind of "evolutionary" algorithm you might be tempted to write (for example, the Bayesian Optimization Algorithm of Pelikan, Goldberg, and Cantú-Paz [2000]). However, if you expect to be able to take an evolutionary computation and have it output an organism on the order of, say, a spider, you will be vastly disappointed. It took roughly a billion years after the start of life for complex cells to arise. Genetic algorithms can design interesting radio antennas, analogous perhaps to a particular chemical enzyme. But even with their hundredfold speedups, modern genetic algorithms seem to be using vastly too little brute force to make it out of the RNA world, let alone reach

101. If intelligence explosion microeconomics tells us that algorithmic advantages are large compared to hardware, then we care most about "nice" projects having the smartest researchers. If hardware advantages are large compared to plausible variance in researcher intelligence, this makes us care more about "nice" projects having the most access to computing resources.

102. Humans count as human-equivalent intelligences.

the Cambrian explosion. To design a spider-equivalent brain would be far beyond the reach of the cumulative optimization power of current evolutionary algorithms running on current hardware for reasonable periods of time.

On the other side of the spectrum, human engineers quite often beat natural selection in particular capacities, even though human engineers have been around for only a tiny fraction of the time. (Wheel beats cheetah, skyscraper beats redwood tree, Saturn V beats falcon, etc.) It seems quite plausible that human engineers, working for an amount of time (or even depth of serial causality) that was small compared to the total number of evolutionary generations, could successfully create human-equivalent intelligence.

However, current AI algorithms fall far short of this level of . . . let's call it "taking advantage of the regularity of the search space," although that's only one possible story about human intelligence. Even branching out into all the fields of AI that try to automatically design small systems, it seems clear that automated design currently falls very far short of human design.

Neither current AI algorithms running on current hardware nor human engineers working on AI for sixty years or so have yet sparked a FOOM. We know two combinations of "algorithm intelligence + amount of search" that haven't output enough cumulative optimization power to spark a FOOM.

But this allows a great deal of room for the possibility that an AI significantly more "efficient" than natural selection, while significantly less "intelligent" than human computer scientists, could start going FOOM. Perhaps the AI would make *less intelligent* optimizations than human computer scientists, but it would make *many more* such optimizations. And the AI would search many fewer individual points in design space than natural selection searched organisms, but traverse the search space *more efficiently* than natural selection.

And, unlike either natural selection or humans, each improvement that the AI found could be immediately reinvested in its future searches. After natural selection built *Homo erectus*, it was not then using *Homo erectus*-level intelligence to consider future DNA modifications. So it might not take very much more intelligence than natural selection for an AI to first build something significantly better than itself, which would then deploy more intelligence to building future successors.

In my present state of knowledge I lack strong information to *not* worry about random AI designs crossing any point on the frontier of "more points searched than any past algorithm of equal or greater intelligence (including human computer scientists), and more intelligence than any past algorithm which has searched an equal number of cases (including natural selection)." This frontier is advanced all the time and no FOOM has yet occurred, so, by Laplace's Rule of Succession or similar ignorance priors, we should assign much less than 50% probability that the next crossing goes FOOM. On

the other hand we should assign a much higher chance that *some* crossing of the frontier of "efficiency cross computation" or "intelligence cross brute force" starts an intelligence explosion at some point in the next $N$ decades.

Our knowledge so far also holds room for the possibility that, without unaffordably vast amounts of computation, semi-intelligent optimizations *cannot* reinvest and cumulate up to human-equivalent intelligence—any more than you can get a FOOM by repeatedly running an optimizing compiler over itself. The theory here is that mice would have a hard time doing better than chance at modifying mice. In this class of scenarios, for any reasonable amount of computation which research projects can afford (even after taking Moore's Law into account), you can't make an AI that builds better AIs than any human computer scientist until that AI is smart enough to actually do computer science. In this regime of possibility, human computer scientists must keep developing their own improvements to the AI until that AI reaches the point of being able to do human-competitive computer science, because until then the AI is not capable of doing very much pushing on its own.[103]

Conversely, to upper-bound the FOOM-starting level, consider the AI equivalent of John von Neumann exploring computer science to greater serial depth and parallel width than previous AI designers ever managed. One would expect this AI to spark an intelligence explosion if it can happen at all. In this case we are going beyond the frontier of the number of optimizations *and* the quality of optimizations for humans, so if this AI can't build something better than itself, neither can humans. The "fast parallel von Neumann" seems like a reasonable pragmatic upper bound on how smart a machine intelligence could be without being able to access an intelligence explosion, or how smart it could be before the intelligence explosion entered a prompt-supercritical mode, assuming this to be possible at all. As it's unlikely for true values to exactly hit upper bounds, I would guess that the intelligence explosion would start well before then.

Relative to my current state of great uncertainty, my median estimate would be somewhere in the middle: that it takes much more than an improved optimizing compiler or improved genetic algorithm, but significantly less than a fast parallel von Neumann,

103. "Nice" AI proposals are likely to *deliberately* look like this scenario, because in Friendly AI we may want to do things like have the AI prove a self-modification correct with respect to a criterion of action—have the AI hold itself to a high standard of self-understanding so that it can change itself in ways which preserve important qualities of its design. This probably implies a large added delay in when a "nice" project can allow its AI to do certain kinds of self-improvement, a significant handicap over less restrained competitors even if the project otherwise has more hardware or smarter researchers. (Though to the extent that you can "sanitize" suggestions or show that a class of improvements can't cause *catastrophic* errors, a Friendly AI under development may be able to wield significant self-improvements even without being able to do computer science.)

to spark an intelligence explosion (in a non-Friendly AI project; a Friendly AI project deliberately requires extra computer science ability in the AI before it is allowed to self-modify). This distribution is based mostly on prior ignorance, but the range seems wide and so the subranges close to the endpoints should be relatively narrow.

All of this range falls well short of what I. J. Good defined as "ultraintelligence." An AI which is merely as good as a fast parallel von Neumann at building AIs need not far surpass humans in all intellectual activities of every sort. For example, it might be very good at computer science while not yet being very good at charismatic manipulation of humans. I. J. Good focused on an assumption that seems far more than sufficient to yield his conclusion of the intelligence explosion, and this unfortunately may be distracting relative to much weaker assumptions that would probably suffice.

### 3.11. Returns on Unknown Unknowns

Molecular nanotechnology is a fairly recent concept and nineteenth-century humans didn't see it coming. There is an important albeit dangerous analogy which says that the twenty-first century can do magic relative to the eleventh century, and yet a thousand years isn't very much time; that to chimpanzees humans are just plain incomprehensible, yet our brain designs aren't even all that different; and that we should therefore assign significant probability that returns on increased speed (serial time, causal depth, more of that distance which separates the twenty-first and eleventh centuries of human history) or improved brain algorithms (more of that which separates hominids from chimpanzees) will end up delivering *damn near anything* in terms of capability.

This may even include capabilities that violate what we currently believe to be the laws of physics, since we may not know all the relevant laws. Of course, just because our standard model of physics might be wrong somewhere, we cannot conclude that any particular error is probable. And new discoveries need not deliver positive news; modern-day physics implies many restrictions the nineteenth century didn't know about, like the speed-of-light limit. Nonetheless, a rational agency will selectively seek out *useful* physical possibilities we don't know about; it will deliberately exploit any laws we do not know. It is not supernaturalism to suspect, in full generality, that future capabilities may somewhere exceed what the twenty-first-century Standard Model implies to be an upper bound.

An important caveat is that if faster-than-light travel is possible by any means whatsoever, the Great Silence/Fermi Paradox ("Where are they?") becomes much harder to explain. This gives us some reason to believe that nobody will ever discover any form of "magic" that enables FTL travel (unless it requires an FTL receiver that must itself travel at slower-than-light speeds). More generally, it gives us a further reason to doubt any future magic in the form of "your physicists didn't know about X, and therefore it is

possible to do Y" that would give many agencies an opportunity to do Y in an observable fashion. We have further reason in addition to our confidence in modern-day physics to believe that time travel is not possible (at least no form of time travel which lets you travel back to before the time machine was built), and that there is no tiny loophole anywhere in reality which even a superintelligence could exploit to enable this, since our present world is not full of time travelers.

More generally, the fact that a rational agency will systematically and selectively seek out previously unknown opportunities for unusually high returns on investment says that the expectation of unknown unknowns should generally drive expected returns upward when dealing with something smarter than us. The true laws of physics might also imply exceptionally bad investment possibilities—maybe even investments worse than the eleventh century would have imagined possible, like a derivative contract that costs only a penny but can lose a quadrillion dollars—but a superintelligence will not be especially interested in those. Unknown unknowns add generic variance, but rational agencies will select on that variance in a positive direction.

From my perspective, the possibility of "returns on unknown unknowns," "returns on magic," or "returns on the superintelligence being smarter than I am and thinking of possibilities I just didn't see coming" mainly tells me that (1) intelligence explosions might go FOOM faster than I expect, (2) trying to bound the real-world capability of an agency *smarter than you are* is unreliable in a fundamental sense, and (3) we probably only get one chance to build something smarter than us that is not uncaring with respect to the properties of the future we care about. But I already believed all that; so, from my perspective, considering the possibility of unknown unknown returns adds little further marginal advice.

Someone else with other background beliefs might propose a wholly different policy whose desirability, given their other beliefs, would hinge mainly on the absence of such unknown unknowns—in other words, it would be a policy whose workability rested on the policy proposer's ability to have successfully bounded the space of opportunities of some smarter-than-human agency. This would result in a rationally unpleasant sort of situation, in the sense that the "argument from unknown unknown returns" seems like it ought to be impossible to defeat, and for an argument to be impossible to defeat means that it is insensitive to reality.[104] I am tempted to say at this point, "Thankfully, that is

104. Indeed, I write these very words in the weary anticipation that somebody is going to claim that the whole AI-go-FOOM thesis, since it could be carried by unknown unknown returns, is actually undefeatable because the argument from magic is undefeatable, and therefore the hard takeoff thesis cannot be defeated by any amount of argument, and therefore belief in it is insensitive to reality, and therefore it is false. I gloomily foretell that pointing out that the whole argument is supposed to carry

not my concern, since my policy proposals are already meant to be optimal replies in the case that a superintelligence can think of something I haven't." But, despite temptation, this brush-off seems inadequately sympathetic to the other side of the debate. And I am not properly sure what sort of procedure ought to be put in place for arguing about the possibility of "returns on unknown unknowns" such that, in a world where there were in fact no significant returns on unknown unknowns, you would be able to figure out with high probability that there were no unknown unknown returns, and plan accordingly.

I do think that proposals which rely on bounding smarter-than-human capacities may reflect a lack of proper appreciation and respect for the notion of something that is *really actually smarter than you*. But it is also not true that the prospect of unknown unknowns means we should assign probability one to a being marginally smarter than human taking over the universe in five seconds, and it is not clear what our actual probability distribution should be over lesser "impossibilities." It is not coincidence that I picked my policy proposal so as not to be highly sensitive to that estimate.

## 4. Three Steps Toward Formality

Lucio Russo (2004), in a book arguing that science was invented two millennia ago and then forgotten, defines an exact science as a body of theoretical postulates whose consequences can be arrived at by unambiguous deduction, which deductive consequences can then be further related to objects in the real world. For instance, by this definition, Euclidean geometry can be viewed as one of the earliest exact sciences, since it proceeds from postulates but also tells us what to expect when we measure the three angles of a real-world triangle.

Broadly speaking, to the degree that a theory is formal, it is possible to say what the theory predicts without argument, even if we are still arguing about whether the theory is actually true. In some cases a theory may be laid out in seemingly formal axioms, and yet its relation to experience—to directly observable facts—may have sufficient flex that people are still arguing over whether or not an agreed-on formal prediction has actually come true.[105] This is often the case in economics: there are many formally specified models of macroeconomics, and yet their relation to experience is ambiguous enough that it's hard to tell which ones, if any, are approximately true.

---

without unknown unknowns, hence its appearance in the final subsection, is not going to have any effect on the repetition of this wonderful counterargument.

105. Another edge case is a formally exact theory whose precise predictions we lack the computing power to calculate, causing people to argue over the deductive consequences of the theory even though the theory's axioms have been fully specified.

What is the point of formality? One answer would be that by making a theory formal, we can compute exact predictions that we couldn't calculate using an intuition in the back of our minds. On a good day, these exact predictions may be unambiguously relatable to experience, and on a truly wonderful day the predictions actually come true.

But this is not the only possible reason why formality is helpful. To make the consequences of a theory subject to unambiguous deduction—even when there is then some further argument over how to relate these consequences to experience—we have to make the machinery of the theory explicit; we have to move it out of the back of our minds and write it out on paper, where it can then be subject to greater scrutiny. This is probably where we will find most of the benefit from trying to analyze the intelligence explosion more formally—it will expose the required internal machinery of arguments previously made informally. It might also tell us startling consequences of propositions we previously said were highly plausible, which we would overlook if we held the whole theory inside our intuitive minds.

With that said, I would suggest approaching the general problem of formalizing previously informal stances on the intelligence explosion as follows:

1. Translate stances into microfoundational hypotheses about growth curves—quantitative functions relating cumulative investment and output. Different stances may have different notions of "investment" and "output," and different notions of how growth curves feed into each other. We want elementary possibilities to be specified with sufficient rigor that their consequences are formal deductions rather than human judgments: in the possibility that X goes as the exponential of Y, then, supposing Y already quantified, the alleged quantity of X should follow as a matter of calculation rather than judgment.

2. Explicitly specify how any particular stance claims that (combinations of) growth curves should allegedly relate to historical observations or other known facts. Quantify the relevant historical observations in a format that can be directly compared to the formal possibilities of a theory, making it possible to formalize a stance's claim that some possibilities in a range are falsified.

3. Make explicit any further assumptions of the stance about the regularity or irregularity (or prior probability) of elementary possibilities. Make explicit any coherence assumptions of a stance about how different possibilities probably constrain each other (curve A should be under curve B, or should have the same shape as curve C).[106]

106. In a Bayesian sense, this corresponds to putting nonindependent joint or conditional prior probabilities over multiple curves.

In the step about relating historical experience to the possibilities of the theory, allowing falsification or updating is importantly not the same as curve-fitting—it's not like trying to come up with a single curve that "best" fits the course of hominid evolution or some such. Hypothesizing that we know a single, exact curve seems like it should be overrunning the state of our knowledge in many cases; for example, we shouldn't pretend to know *exactly* how difficult it was for natural selection to go from *Homo erectus* to *Homo sapiens*. To get back a prediction with appropriately wide credible intervals—a prediction that accurately represents a state of uncertainty—there should be some space of regular curves in the model space, with combinations of those curves related to particular historical phenomena. In principle, we would then falsify the combinations that fail to match observed history, and integrate (or sample) over what's left to arrive at a prediction.

Some widely known positions on the intelligence explosion do rely on tightly fitting a curve (e.g., Moore's Law). This is not completely absurd because some historical curves have in fact been highly regular (e.g., Moore's Law). By passing to Bayesian updating instead of just falsification, we could promote parts of the model space that *narrowly* predict an observed curve—parts of the model space which concentrated more of their probability mass into predicting that exact outcome. This would expose assumptions about likelihood functions and make more visible whether it's reasonable or unreasonable to suppose that a curve is precise; if we do a Bayesian update on the past, do we get narrow predictions for the future? What do we need to assume to get narrow predictions for the future? How steady has Moore's Law actually been for the past?—because if our modeling technique can't produce even that much steadiness, and produces wide credible intervals going off in all directions, then we're not updating hard enough or we have overly ignorant priors.

Step One would be to separately carry out this process on one or more current stances and speakers, so as to reveal and quantify the formal assumptions underlying their arguments. At the end of Step One, you would be able to say, "This is a model space that looks like what Speaker X was talking about; these are the growth curves or combinations of growth curves that X considers falsified by these historical experiences, or that X gives strong Bayesian updates based on their narrow predictions of historical experiences; this is what X thinks about how these possibilities are constrained to be coherent with each other; and this is what X thinks is the resulting prediction made over the intelligence explosion by the nonfalsified, coherent parts of the model space."

Step One of formalization roughly corresponds to seeing if there's *any* set of curves by which a speaker's argument could make sense; making explicit the occasions where someone else has argued that possibilities are excluded by past experience; and exposing any suspicious irregularities in the curves being postulated. Step One wouldn't yield

definitive answers about the intelligence explosion, but should force assumptions to be more clearly stated, potentially expose some absurdities, show what else a set of assumptions implies, etc. Mostly, Step One is about explicitizing stances on the intelligence explosion, with each stance considered individually and in isolation.

Step Two would be to try to have a common, integrated model of multiple stances formalized in Step One—a model that included many different possible kinds of growth curves, some of which might be (in some views) already falsified by historical observations—a common pool of building blocks that could be selected and snapped together to produce the individual formalizations from Step One. The main products of Step Two would be (a) a systematic common format for talking about plausible growth curves and (b) a large table of which assumptions yield which outcomes (allegedly, according to the compiler of the table) and which historical observations various arguments allege to pose problems for those assumptions. I would consider this step to be about making explicit the *comparison* between theories: exposing arguable irregularities that exist in one stance but not another and giving readers a better position from which to evaluate supposed better matches versus simpler hypotheses. Step Two should not yet try to take strong positions on the relative plausibility of arguments, nor to yield definitive predictions about the intelligence explosion. Rather, the goal is to make comparisons between stances more formal and more modular, without leaving out any important aspects of the informal arguments—to formalize the conflicts between stances in a unified representation.

Step Three would be the much more ambitious project of coming up with an allegedly uniquely correct description of our state of uncertain belief about the intelligence explosion:

- Formalize a model space broad enough to probably contain something like reality, with credible hope of containing a point hypothesis in its space that would well fit, if not exactly represent, whatever causal process actually turns out to underlie the intelligence explosion. That is, the model space would not be so narrow that, if the real-world growth curve were actually hyperbolic up to its upper bound, we would have to kick ourselves afterward for having no combinations of assumptions in the model that could possibly yield a hyperbolic curve.[107]
- Over this model space, weight prior probability by simplicity and regularity.

107. In other words, the goal would be to avoid errors of the class "nothing like the reality was in your hypothesis space at all." There are many important theorems of Bayesian probability that do not apply when nothing like reality is in your hypothesis space.

- Relate combinations of causal hypotheses to observed history and do Bayesian updates.
- Sample the updated model space to get a probability distribution over the answers to any query we care to ask about the intelligence explosion.
- Tweak bits of the model to get a sensitivity analysis of how much the answers tend to vary when you model things slightly differently, delete parts of the model to see how well the coherence assumptions can predict the deleted parts from the remaining parts, etc.

If Step Three is done wisely—with the priors reflecting an appropriate breadth of uncertainty—and doesn't entirely founder on the basic difficulties of formal statistical learning when data is scarce, then I would expect any such formalization to yield mostly qualitative yes-or-no answers about a rare handful of answerable questions, rather than yielding narrow credible intervals about exactly how the internal processes of the intelligence explosion will run. A handful of yeses and nos is about the level of advance prediction that I think a reasonably achievable grasp on the subject *should* allow—we *shouldn't* know most things about intelligence explosions this far in advance of observing one—we should just have a few rare cases of questions that have highly probable if crude answers. I think that one such answer is "AI go FOOM? Yes! AI go FOOM!" but I make no pretense of being able to state that it will proceed at a rate of 120,000 nanofooms per second.

Even at that level, covering the model space, producing a reasonable simplicity weighting, correctly hooking up historical experiences to allow falsification and updating, and getting back the rational predictions would be a rather ambitious endeavor that would be easy to get wrong. Nonetheless, I think that Step Three describes in principle what the ideal Bayesian answer would be, given our current collection of observations. In other words, the reason I endorse an AI-go-FOOM answer is that I think that our historical experiences falsify most regular growth curves over cognitive investments that wouldn't produce a FOOM.

Academic disputes are usually not definitively settled once somebody advances to the stage of producing a simulation. It's worth noting that macroeconomists are still arguing over, for example, whether inflation or NGDP should be stabilized to maximize real growth. On the other hand, macroeconomists usually want more precise answers than we could reasonably demand from predictions about the intelligence explosion. If you'll settle for model predictions like, "Er, maybe inflation ought to increase rather than decrease when banks make noticeably more loans, *ceteris paribus*?" then it might be more reasonable to expect definitive answers, compared to asking whether inflation will be more or less than 2.3%. But even if you tried to build *the* Step Three model, it

might still be a bit naive to think that you would really get *the* answers back out, let alone expect that everyone else would trust your model.

In my case, I think how much I trusted a Step Three model would depend a lot on how well its arguments simplified, while still yielding the same net predictions and managing not to be falsified by history. I trust complicated arguments much more when they have simple versions that give mostly the same answers; I would trust my arguments about growth curves less if there weren't also the simpler version, "Smart minds build even smarter minds." If the model told me something I hadn't expected, but I could translate the same argument back into simpler language and the model produced similar results even when given a few cross-validational shoves, I'd probably believe it.

Regardless, we can legitimately hope that finishing Step One, going on to Step Two, and pushing toward Step Three will yield interesting results, even if Step Three is never completed or is completed several different ways.[108] The main point of formality isn't that it gives you final and authoritative answers, but that it sometimes turns up points you wouldn't have found without trying to make things explicit.

## 5. Expected Information Value: What We Want to Know versus What We Can Probably Figure Out

There tend to be mismatches between what we want to know about the intelligence explosion, and what we can reasonably hope to figure out.

For example, everyone at the Machine Intelligence Research Institute (MIRI) would love to know how much time remained until an intelligence explosion would probably be produced by general progress in the field of AI. It would be extremely useful knowledge from a policy perspective, and if you could time it down to the exact year, you could run up lots of credit card debt just beforehand.[109] But—unlike a number of other futurists—we don't see how we could reasonably obtain strong information about this question.

Hans Moravec, one of the first major names to predict strong AI using Moore's Law, spent much of his (1988) book *Mind Children* trying to convince readers of the incredible proposition that Moore's Law could actually go on continuing and continuing and continuing until it produced supercomputers that could do—gasp!—a hundred teraflops. Which was enough to "equal the computing power of the human brain," as Moravec had calculated that equivalency in some detail using what was then known about the visual cortex and how hard that part was to simulate. We got the supercomputers that

108. "A man with one watch knows what time it is; a man with two watches is never sure."

109. Yes, that is a joke.

Moravec thought were necessary in 2008, several years earlier than Moravec's prediction; but, as it turned out, the way reality works is not that the universe checks whether your supercomputer is large enough and then switches on its consciousness.[110] Even if it were a matter of hardware rather than mostly software, the threshold level of "required hardware" would be far more uncertain than Moore's Law, and a predictable number times an unpredictable number is an unpredictable number.

So, although there is an extremely high value of information about default AI timelines, our expectation that formal modeling can update our beliefs about this quantity is low. We would mostly expect modeling to formally tell us, "Since this quantity depends conjunctively on many variables you're uncertain about, you are very uncertain about this quantity." It would make some sense to poke and prod at the model to see if it had something unexpected to say—but I'd mostly expect that we can't, in fact, produce tight credible intervals over default AI arrival timelines given our state of knowledge, since this number sensitively depends on many different things we don't know. Hence my strong statement of normative uncertainty: "I don't know which decade and you don't know either!"

(Even this kind of "I don't know" still has to correspond to some probability distribution over decades, just not a tight distribution. I'm currently trying to sort out with Carl Shulman why my median is forty-five years in advance of his median. Neither of us thinks we can time it down to the decade—we have very broad credible intervals in both cases—but the discrepancy between our "I don't knows" is too large to ignore.)

Some important questions on which policy depends—questions I would want information about, where it seems there's a reasonable chance that new information might be produced, with direct links to policy—are as follows:

- How likely is an intelligence explosion to be triggered by a relatively dumber-than-human AI that can self-modify more easily than us? (This is policy relevant because it tells us how early to worry. I don't see particularly how this information could be obtained, but I also don't see a strong argument saying that we have to be ignorant of it.)
- What is the slope of the self-improvement curve in the near vicinity of roughly human-level intelligence? Are we confident that it'll be "going like gangbusters" at that point and not slowing down until later? Or are there plausible and probable scenarios in which human-level intelligence was itself achieved as the result of a

110. See also *The Moon is a Harsh Mistress* (Heinlein 1966) and numerous other SF stories that made the same assumption (big computer = intelligence, or complex computer = consciousness) as a cheap way to throw an AI into the story. A different SF story, *Death in the Promised Land* (Cadigan 1995), compared this to the ancient theory that dirty shirts and straw would spontaneously generate mice.

self-improvement curve that had already used up all low-hanging fruits to that point? Or human researchers pushed the AI to that level and it hasn't self-improved much as yet? (This is policy relevant because it determines whether there's any substantial chance of the world having time to react after AGI appears in such blatant form that people actually notice.)

- Are we likely to see a relatively smooth or relatively "jerky" growth curve in early stages of an intelligence explosion? (Policy relevant because sufficiently smooth growth implies that we can be less nervous about promising systems that are currently growing slowly, keeping in mind that a heap of uranium bricks is insufficiently smooth for policy purposes despite its physically continuous behavior.)

Another class of questions which are, in pragmatic practice, worth analyzing, are those on which a more formal argument might be more accessible to outside academics. For example, I hope that formally modeling returns on cognitive reinvestment, and constraining those curves by historical observation, can predict "AI go FOOM" in a way that's more approachable to newcomers to the field.[111] But I would derive little personal benefit from being formally told, "AI go FOOM," even with high confidence, because that was something I already assigned high probability on the basis of "informal" arguments, so I wouldn't shift policies. Only expected belief updates that promise to yield policy shifts can produce expected value of information.

(In the case where I'm just plain wrong about FOOM for reasons exposed to me by formal modeling, this produces a drastic policy shift and hence extremely high value of information. But this result would be, at least to me, surprising; I'd mostly expect to get back an answer of "AI go FOOM" or, more probably for early modeling attempts, "Dunno.")

But pragmatically speaking, if we can well-formalize the model space and it does yield a prediction, this would be a very nice thing to have around properly written up. So, pragmatically, this particular question is worth time to address.

Some other questions where I confess to already having formed an opinion, but for which a more formal argument would be valuable, and for which a surprising weakness would of course be even more valuable:

111. Of course I would try to invoke the discipline of Anna Salamon to become curious if an *a priori* trustworthy-seeming modeling attempt came back and said, "AI definitely not go FOOM." Realistically, I probably wouldn't be able to stop myself from expecting to find a problem in the model. But I'd also try not to impose higher burdens of proof, try to look equally skeptically at parts that seemed *congruent* with my prior beliefs, and generally not toss new evidence out the window or be "that guy" who can't change his mind about anything. And others at MIRI and interested outsiders would have less strong prior beliefs.

- Is human intelligence the limit of the possible? Is there a "General Intelligence Theorem" à la Greg Egan which says that nothing qualitatively smarter than a human can exist?
- Does I. J. Good's original argument for the intelligence explosion carry? Will there be a historically unprecedented upsurge in intelligence that gets to the level of strong superintelligence before running out of steam?
- Will the intelligence explosion be relatively local or relatively global? Is this something that happens inside one intelligence, or is it a grand function of the total world economy? Should we expect to see a civilization that grew out of many AI projects that traded data with each other, with no single AI becoming stronger than the others; or should we expect to see an AI singleton?[112]

Policy-relevant questions that I wish I could get data about, but for which I don't think strong data is likely to be available, or about which microeconomic methodology doesn't seem to have much to say:

- How much time remains before general progress in the field of AI is likely to generate a successful AGI project?
- How valuable are smarter researchers to an AI project, versus a thousand times as much computing power?
- What's the top warning sign that an individual AI project is about to go FOOM? What do AIs look like just before they go FOOM?

More generally, for every interesting-sounding proposition X, we should be interested in *any* strong conclusions that an investigation claims to yield, such as:

- Definitely not-X, because a model with X strongly implies growth curves that look like they would violate our previous historical experience, or curves that would have to undergo specific unexplained irregularities as soon as they're out of regimes corresponding to parts we've already observed. (The sort of verdict you might expect for the sometimes-proffered scenario that "AI will advance to the human level and then halt.")
- Definitely X, because nearly all causal models that we invented and fit to historical experience, and then adapted to query what would happen for self-improving AI,

112. Here I'm somewhat uncertain about the "natural" course of events, but I feel less personal curiosity because I will still be trying to build a Friendly AI that does a local FOOM even if this is a moderately "unnatural" outcome.

yielded X without further tweaking throughout almost all their credible intervals. (This is how I think we should formalize the informal argument put forth for why we should expect AI to undergo an intelligence explosion, given that natural selection didn't seem to run into hardware or software barriers over the course of hominid evolution, etc.)

- We definitely don't know whether X or not-X, and nobody else could possibly know either. All plausible models show that X varies strongly with Y and Z, and there's no reasonable way anyone could know Y, and even if they did, they still wouldn't know Z.[113] (The sort of formal analysis we might plausibly expect for "Nobody knows the timeline to strong AI.") Therefore, a rational agent should assign probabilities using this highly ignorant prior over wide credible intervals, and should act accordingly by planning for and preparing for multiple possible outcomes. (Note that in some cases this itself equates to an antiprediction, a strong ruling against a "privileged" possibility that occupies only a narrow range of possibility space. If you definitely can't predict something on a wide logarithmic scale, then as a matter of subjective probability it is unlikely to be within a factor of three of some sweet spot, and scenarios which require the sweet spot are *a priori* improbable.)

## 6. Intelligence Explosion Microeconomics: An Open Problem

My proposed project of intelligence explosion microeconomics can be summarized as follows:

> Formalize stances on the intelligence explosion in terms of microfoundational growth curves and their interaction, make explicit how past observations allegedly constrain those possibilities, and formally predict future outcomes based on such updates.

This only reflects one particular idea about methodology, and more generally the open problem could be posed thus:

> Systematically answer the question, "What do we think we know and how do we think we know it?" with respect to growth rates of cognitive reinvestment.

Competently undertaking the entire project up to Step Three would probably be a PhD-thesis-sized project, or even a multiresearcher project requiring serious funding. Step One investigations might be doable as smaller-scale projects, but would still be difficult.

113. Katja Grace observes abstractly that X might still (be known to) correlate strongly with some observable W, which is a fair point.

MIRI is highly interested in trustworthy progress on this question that offers to resolve our actual internal debates and policy issues, but this would require a high standard of work (the formal model has to be competitive with highly developed informal models) and considerable trust that the researcher wasn't entering with strong biases in any particular direction (motivated cognition), including any biases in favor of making the results come out neutral (motivated neutrality) or uncertain (motivated uncertainty). We would only sponsor work on this project if we expected a sufficiently high ratio of "hope of getting real answers we didn't already know/cost of funding the project."

Potential investigators should have:

- Some amount of prior experience with mathematical economics. Failing that, at least some knowledge of standard econ-with-math, plus being able to formulate and solve differential equations.
- Enough statistical prediction/machine learning experience to know what happens when you try to fit a model with lots of parameters without doing regularization and cross-validation.
- A demonstrably strong intuitive sense for what all those fancy equations *mean*: being the sort of person who asks, "But if it always takes exponentially larger brains to get linear increases in intelligence, then how do you square that with human brain sizes versus chimpanzee brain sizes?"
- Enough familiarity with the cognitive science literature and/or basic epistemic skills that you are explicitly aware of and on guard against motivated credulity, motivated skepticism, packing and unpacking, expert overconfidence, the conjunction fallacy, the history of Millikan's oil-drop experiment, etc. Ideally (though this is not required) you will be familiar with some locally grown concepts like motivated stopping and continuation, motivated neutrality, motivated uncertainty, etc.
- Being demonstrably able to write up results for publication. We care significantly about making results accessible to the general public, as well as about knowing them ourselves.
- Prior familiarity with the literature on the intelligence explosion, including our own literature, is *not* on this list. Such acquaintance can be obtained afterward by skimming the (few) previous informal debates and directly talking to the (few) major players to confirm your interpretations of their stances.

This may sound like a high bar, and a lot of work—but we're talking about what it would take to do the canonical growth-rate analysis of a purported future phenomenon, I. J. Good's intelligence explosion, which if real is probably the most important phenomenon in the history of Earth-originating intelligent life. If there are in fact no

aliens within the range of our telescopes, the intelligence explosion will plausibly be the most important event determining the future of the visible universe. Trustworthy information about any predictable aspect of the intelligence explosion is highly valuable and important.

To foster high-quality research on intelligence explosion microeconomics, MIRI has set up a private mailing list for qualified researchers. MIRI will publish its own research on the subject to this mailing list first, as may other researchers. If you would like to apply to join this mailing list, contact MIRI for instructions (admin@intelligence.org).

## Acknowledgments

My thanks to Katja Grace for her research assistance, and to Carl Shulman, Paul Christiano, Luke Muehlhauser, Katja Grace, Kaj Sotala, Robin Hanson, Nick Bostrom, Moshe Looks, and Benjamin Noble for their helpful feedback.

## References

Armstrong, Stuart. Forthcoming. "General Purpose Intelligence: Arguing the Orthogonality Thesis." *Analysis and Metaphysics*. Preprint at http://lesswrong.com/lw/cej/general_purpose_intelligence_arguing_the/.

Barbour, Julian. 1999. *The End of Time: The Next Revolution in Physics*. 1st ed. New York: Oxford University Press.

Baum, Eric B. 2004. *What Is Thought?* Bradford Books. Cambridge, MA: MIT Press.

Bickerton, Derek. 2009. *Adam's Tongue: How Humans Made Language, How Language Made Humans*. New York: Hill & Wang.

Blair, Clay, Jr. 1957. "Passing of a Great Mind: John von Neumann, a Brilliant, Jovial Mathematician, was a Prodigious Servant of Science and His Country." *Life*, February 25, 89–104. `http://books.google.ca/books?id=rEEEAAAAMBAJ&pg=PA89`.

Bostrom, Nick. 2006. "What is a Singleton?" *Linguistic and Philosophical Investigations* 5 (2): 48–54.

———. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." In "Theory and Philosophy of AI," edited by Vincent C. Müller. Special issue, *Minds and Machines* 22 (2): 71–85. doi:`10.1007/s11023-012-9281-3`.

Bringsjord, Selmer. 2012. "Belief in the Singularity is Logically Brittle." *Journal of Consciousness Studies* 19 (7-8): 14–20. `http://ingentaconnect.com/content/imp/jcs/2012/00000019/F0020007/art00002`.

Cadigan, Pat. 1995. "Death in the Promised Land." *Omni Online*, March.

Calvin, William H. 2004. *A Brief History of the Mind: From Apes to Intellect and Beyond*. New York: Oxford University Press.

Chalmers, David John. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (9–10): 7–65. `http://www.ingentaconnect.com/content/imp/jcs/2010/00000017/f0020009/art00001`.

———. 2012. "The Singularity: A Reply to Commentators." *Journal of Consciousness Studies* 19 (7-8): 141–167. `http://ingentaconnect.com/content/imp/jcs/2012/00000019/F0020007/art00014`.

Chalmers, David John, Robert M. French, and Douglas R. Hofstadter. 1992. "High-Level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology." *Journal of Experimental and Theoretical Artificial Intelligence* 4 (3): 185–211. doi:`10.1080/09528139208953747`.

Clark, Gregory. 2007. *A Farewell to Alms: A Brief Economic History of the World*. 1st ed. Princeton, NJ: Princeton University Press.

Cole, David. 2013. "The Chinese Room Argument." In *The Stanford Encyclopedia of Philosophy,* Spring 2013, edited by Edward N. Zalta. Stanford University. http://plato.stanford.edu/archives/spr2013/entries/chinese-room/.

Cowen, Tyler. 2011. *The Great Stagnation: How America Ate All the Low-Hanging Fruit of Modern History, Got Sick, and Will (Eventually) Feel Better.* New York: Dutton.

Deacon, Terrence W. 1997. *The Symbolic Species: The Co-evolution of Language and the Brain.* New York: W. W. Norton.

Eden, Amnon, Johnny Søraker, James H. Moor, and Eric Steinhart, eds. 2012. *Singularity Hypotheses: A Scientific and Philosophical Assessment.* The Frontiers Collection. Berlin: Springer.

Egan, Greg. 2002. *Schild's Ladder.* New York: Eos.

Falkenhainer, Brian, and Kenneth D. Forbus. 1990. "The Structure-Mapping Engine: Algorithm and Examples." *Artificial Intelligence* 41 (1): 1–63. doi:10.1016/0004-3702(89)90077-5.

Feldman, J. A., and Dana H. Ballard. 1982. "Connectionist Models and Their Properties." *Cognitive Science* 6 (3): 205–254. doi:10.1207/s15516709cog0603_1.

Frankena, William K. 1973. *Ethics.* 2nd ed. Foundations of Philosophy Series. Englewood Cliffs, NJ: Prentice-Hall.

Freitas, Robert A., Jr. 2000. "Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations." Foresight Institute. April. Accessed July 28, 2013. http://www.foresight.org/nano/Ecophagy.html.

Gibson, William. 1984. *Neuromancer.* 1st ed. New York: Ace.

Good, Irving John. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers,* edited by Franz L. Alt and Morris Rubinoff, 31–88. Vol. 6. New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0.

Hanson, Robin. 1998a. "Economic Growth Given Machine Intelligence." Unpublished manuscript. Accessed May 15, 2013. http://hanson.gmu.edu/aigrow.pdf.

———. 1998b. "Long-Term Growth as a Sequence of Exponential Modes." Unpublished manuscript. Last revised December 2000. http://hanson.gmu.edu/longgrow.pdf.

———. 1998c. "Must Early Life Be Easy? The Rhythm of Major Evolutionary Transitions." Unpublished manuscript, September 23. Accessed August 12, 2012. http://hanson.gmu.edu/hardstep.pdf.

———. 2008a. "Outside View of the Singularity." *Overcoming Bias* (blog), June 20. http://www.overcomingbias.com/2008/06/singularity-out.html.

———. 2008b. "Shared AI Wins." *Overcoming Bias* (blog), December 6. http://www.overcomingbias.com/2008/12/shared-ai-wins.html.

———. 2008c. "Test Near, Apply Far." *Overcoming Bias* (blog), December 3. http://www.overcomingbias.com/2008/12/test-near-apply.html.

———. 2008d. "The Rapacious Hardscrapple Frontier." In *Year Million: Science at the Far Edge of Knowledge,* edited by Damien Broderick, 168–189. New York: Atlas. http://hanson.gmu.edu/hardscra.pdf.

Hawks, John, Eric T. Wang, Gregory M. Cochran, Henry C. Harpending, and Robert K. Moyzis. 2007. "Recent Acceleration of Human Adaptive Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 104 (52): 20753–20758. doi:10.1073/pnas.0707650104.

Heinlein, Robert A. 1966. *The Moon is a Harsh Mistress*. New York: Putnam.

Kahneman, Daniel, and Dan Lovallo. 1993. "Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking." *Management Science* 39 (1): 17–31. doi:10.1287/mnsc.39.1.17.

Kasparov, Garry, and Daniel King. 2000. *Kasparov Against the World: The Story of the Greatest Online Challenge*. New York: KasparovChess Online.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. 1st ed. Chicago: University of Chicago Press.

Kurzweil, Ray. 1999. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. New York: Viking.

Langley, Patrick, Gary Bradshaw, and Jan Zytkow. 1987. *Scientific Discovery: Computational Explorations of the Creative Process*. Cambridge, MA: MIT Press.

Lucas, Robert E., Jr. 1976. "Econometric Policy Evaluations: A Critique." *Carnegie-Rochester Conference Series on Public Policy* 1:19–46. doi:10.1016/S0167-2231(76)80003-6.

Mahoney, Matt. 2010. "A Model for Recursively Self Improving Programs v.3." Unpublished manuscript, December 17. Accessed March 27, 2012. http://mattmahoney.net/rsi.pdf.

McDaniel, Michael A. 2005. "Big-Brained People are Smarter: A Meta-Analysis of the Relationship between In Vivo Brain Volume and Intelligence." *Intelligence* 33 (4): 337–346. doi:10.1016/j.intell.2004.11.005.

Moravec, Hans P. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press.

———. 1999. "Simple Equations for Vinge's Technological Singularity." Unpublished manuscript, February. http://www.frc.ri.cmu.edu/~hpm/project.archive/robot.papers/1999/singularity.html.

Muehlhauser, Luke, and Louie Helm. 2012. "The Singularity and Machine Ethics." In Eden, Søraker, Moor, and Steinhart 2012.

Muehlhauser, Luke, and Anna Salamon. 2012. "Intelligence Explosion: Evidence and Import." In Eden, Søraker, Moor, and Steinhart 2012.

Muehlhauser, Luke, and Chris Williamson. 2013. *Ideal Advisor Theories and Personal CEV*. Berkeley, CA: Machine Intelligence Research Institute. http://intelligence.org/files/IdealAdvisorTheories.pdf.

NSB (National Science Board). 2012. *Science and Engineering Indicators 2012*. NSB 12-01. Arlington, VA: National Science Foundation. http://www.nsf.gov/statistics/seind12/start.htm.

Omohundro, Stephen M. 2008. "The Basic AI Drives." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

Pelikan, Martin, David E. Goldberg, and Erick Cantú-Paz. 2000. "Linkage Problem, Distribution Estimation, and Bayesian Networks." *Evolutionary Computation* 8 (3): 311–340. doi:10.1162/106365600750078808.

Pennachin, Cassio, and Ben Goertzel. 2007. "Contemporary Approaches to Artificial General Intelligence." In *Artificial General Intelligence*, edited by Ben Goertzel and Cassio Pennachin, 1–30. Cognitive Technologies. Berlin: Springer. doi:10.1007/978-3-540-68677-4_1.

Ponce de León, Marcia S., Lubov Golovanova, Vladimir Doronichev, Galina Romanova, Takeru Akazawa, Osamu Kondo, Hajime Ishida, and Christoph P. E. Zollikofer. 2008. "Neanderthal Brain Size at Birth Provides Insights into the Evolution of Human Life History." *Proceedings of the National Academy of Sciences of the United States of America* 105 (37): 13764–13768. doi:10.1073/pnas.0803917105.

Rawls, John. 1971. *A Theory of Justice.* Cambridge, MA: Belknap.

Rhodes, Richard. 1986. *The Making of the Atomic Bomb.* New York: Simon & Schuster.

Rosati, Connie S. 1995. "Persons, Perspectives, and Full Information Accounts of the Good." *Ethics* 105 (2): 296–325. doi:10.1086/293702.

Russo, Lucio. 2004. *The Forgotten Revolution: How Science Was Born in 300 BC and Why It Had to Be Reborn.* Translated by Silvio Levy. New York: Springer.

Sandberg, Anders. 2010. "An Overview of Models of Technological Singularity." Paper presented at the Roadmaps to AGI and the Future of AGI Workshop, Lugano, Switzerland, March 8. http://agi-conf.org/2010/wp-content/uploads/2009/06/agi10singmodels2.pdf.

Shulman, Carl, and Nick Bostrom. 2012. "How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects." *Journal of Consciousness Studies* 19 (7–8): 103–130. http://ingentaconnect.com/content/imp/jcs/2012/00000019/F0020007/art00011.

Silver, Nate. 2012. *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't.* New York: Penguin.

Sternberg, Robert J., and Scott Barry Kaufman, eds. 2011. *The Cambridge Handbook of Intelligence.* Cambridge Handbooks in Psychology. New York: Cambridge University Press.

Tegmark, Max. 2000. "Importance of Quantum Decoherence in Brain Processes." *Physical Review E* 61 (4): 4194–4206. doi:10.1103/PhysRevE.61.4194.

Tuomi, Ilkka. 2002. "The Lives and the Death of Moore's Law." *First Monday* 7 (11). http://firstmonday.org/ojs/index.php/fm/article/view/1000/921.

*Wikipedia.* 2013, s.v. "Lucas Critique." Accessed April 11. http://en.wikipedia.org/w/index.php?title=Lucas_critique&oldid=549911736.

Wiles, Andrew. 1995. "Modular Elliptic Curves and Fermat's Last Theorem." *Annals of Mathematics* 142 (3): 443–551. doi:10.2307/2118559.

Williams, George C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought.* Princeton Science Library. Princeton, NJ: Princeton University Press.

Williams, Roger. 2002. *The Metamorphosis of Prime Intellect.* http://localroger.com/prime-intellect/mopiidx.html.

Yudkowsky, Eliezer. 2007. "Evolutions Are Stupid (But Work Anyway)." *Less Wrong* (blog), November 3. http://lesswrong.com/lw/kt/evolutions_are_stupid_but_work_anyway/.

———. 2008a. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks,* edited by Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.

———. 2008b. "Optimization and the Singularity." *Less Wrong* (blog), June 23. http://lesswrong.com/lw/rk/optimization_and_the_singularity/.

———. 2008c. "Surprised by Brains." *Less Wrong* (blog), November 23. http://lesswrong.com/lw/w4/surprised_by_brains/.

———. 2008d. "The First World Takeover." *Less Wrong* (blog), November 19. http://lesswrong.com/lw/w0/the_first_world_takeover/.

———. 2010. "'Outside View!' as Conversation-Halter." *Less Wrong* (blog), February 24. http://lesswrong.com/lw/1p5/outside_view_as_conversationhalter/.

———. 2011. "Complex Value Systems in Friendly AI." In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings*, edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 388–393. Lecture Notes in Computer Science 6830. Berlin: Springer. doi:10.1007/978-3-642-22887-2_48.