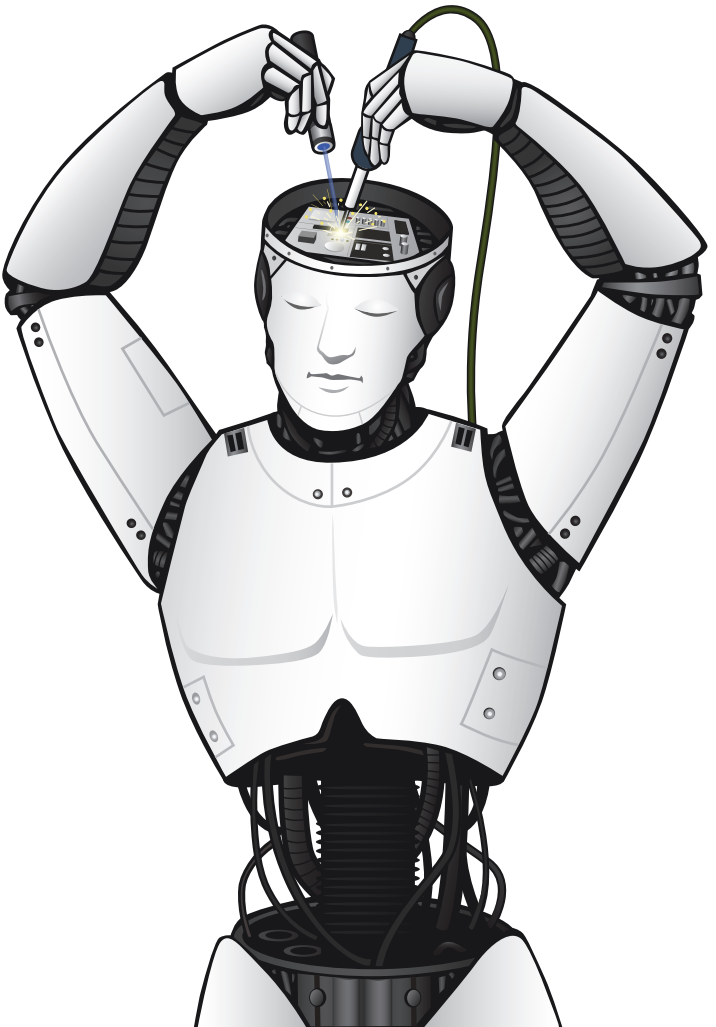
THE HANSON-YUDKOWSKY AI-FOOM DEBATE

ROBIN HANSON AND ELIEZER YUDKOWSKY



The Hanson-Yudkowsky AI-Foom Debate

The Hanson-Yudkowsky AI-Foom Debate

Robin Hanson and Eliezer Yudkowsky

Robin Hanson is an associate professor of economics at George Mason University and a research associate at the Future of Humanity Institute of Oxford University.

Eliezer Yudkowsky is a Research Fellow at the Machine Intelligence Research Institute and is the foremost researcher on Friendly AI and recursive self-improvement.

Published in 2013 by the
Machine Intelligence Research Institute,
Berkeley 94704
United States of America
intelligence.org

ISBN-10: 1939311039
ISBN-13: 978-1-939311-03-0
(PDF)

The Machine Intelligence Research Institute gratefully acknowledges each of the authors for their ideas and contributions toward this important topic. Special thanks to Carl Shulman and James Miller for their guest posts in the debate.

Contents

Foreword

In late 2008, economist Robin Hanson and AI theorist Eliezer Yudkowsky conducted an online debate about the future of artificial intelligence, and in particular about whether generally intelligent AIs will be able to improve their own capabilities very quickly (a.k.a. "foom"). James Miller and Carl Shulman also contributed guest posts to the debate.

The original debate took place in a long series of blog posts, which are collected here. This book also includes a transcript of a 2011 in-person debate between Hanson and Yudkowsky on this subject, a summary of the debate written by Kaj Sotala, and a 2013 technical report on AI takeoff dynamics ("intelligence explosion microeconomics") written by Yudkowsky.

Comments from the authors are included at the end of each chapter, along with a link to the original post. The curious reader is encouraged to use these links to view the original posts and all com-

ments. This book contains minor updates, corrections, and additional citations.

Part I

Prologue

1

Fund UberTool*?*

Robin Hanson

12 November 2008

Some companies specialize in making or servicing tools, and some even specialize in redesigning and inventing tools. All these tool companies use tools themselves. Let us say that tool type A "aids" tool type B if tools of type A are used when improving tools of type B. The aiding graph can have cycles, such as when A aids B aids C aids D aids A.

Such tool aid cycles contribute to progress and growth. Sometimes a set of tool types will stumble into conditions especially favorable for mutual improvement. When the aiding cycles are short and the aiding relations are strong, a set of tools may improve together especially quickly. Such favorable storms of mutual improvement usually run out quickly, however, and in all of human history no more

than three storms have had a large and sustained enough impact to substantially change world economic growth rates.[1]

Imagine you are a venture capitalist reviewing a proposed business plan. *UberTool Corp* has identified a candidate set of mutually aiding tools, and plans to spend millions pushing those tools through a mutual improvement storm. While *UberTool* may sell some minor patents along the way, *UberTool* will keep its main improvements to itself and focus on developing tools that improve the productivity of its team of tool developers.

In fact, *UberTool* thinks that its tool set is so fantastically capable of mutual improvement, and that improved versions of its tools would be so fantastically valuable and broadly applicable, *UberTool* does not plan to stop their closed self-improvement process until they are in a position to suddenly burst out and basically "take over the world." That is, at that point their costs would be so low they could enter and dominate most industries.

Now given such enormous potential gains, even a very tiny probability that *UberTool* could do what they planned might entice you to invest in them. But even so, just what exactly would it take to convince you *UberTool* had even such a tiny chance of achieving such incredible gains?

* * *

Eliezer Yudkowsky

. . . I'll offer my own intuitive answer to the above question: You've got to be doing something that's the same order of Cool as the invention of "animal brains, human brains, farming, and industry." I think this is the wrong list, really; "farming" sets too low a standard. And certainly venture capitalists have a tendency and a motive to exaggerate how neat their projects are.

But if, without exaggeration, you find yourself saying, "Well, that looks like a much larger innovation than farming"—so as to leave some safety margin—then why shouldn't it have at least that large an impact?

However, I would be highly skeptical of an *UberTool Corp* that talked about discounted future cash flows and return on investment. I would be suspicious that they weren't acting the way I would expect someone to act if they really believed in their *UberTool*.

See original post for all comments.

* * *

1. Robin Hanson, "In Innovation, Meta is Max," *Overcoming Bias* (blog), June 15, 2008, http://www.overcomingbias.com/2008/06/meta-is-max---i.html.

2

Engelbart as UberTool*?*

Robin Hanson

13 November 2008

Yesterday I described *UberTool*, an imaginary company planning to push a set of tools through a mutual-improvement process; their team would improve those tools, and then use those improved versions to improve them further, and so on through a rapid burst until they were in a position to basically "take over the world." I asked what it would take to convince you their plan was reasonable, and got lots of thoughtful answers.

Douglas Engelbart is the person I know who came closest to enacting such a *UberTool* plan. His seminal 1962 paper, "Augmenting Human Intellect: A Conceptual Framework," proposed using computers to create such a rapidly improving tool set.[1] He understood not just that computer tools were especially open to mutual improvement, but also a lot about what those tools would look like. Wikipedia:

> [Engelbart] is best known for inventing the computer mouse . . . [and] as a pioneer of human-computer interaction whose team developed hypertext, networked computers, and precursors to GUIs.[2]

Doug led a team who developed a rich set of tools including a working hypertext publishing system. His 1968 "Mother of all Demos" to a thousand computer professionals in San Francisco

> featured the first computer mouse the public had ever seen, as well as introducing interactive text, video conferencing, teleconferencing, email and hypertext [= the web].[3]

Now to his credit, Doug never suggested that his team, even if better funded, might advance so far so fast as to "take over the world." But he did think it could go far (his Bootstrap Institute still pursues his vision), and it is worth pondering just how far it was reasonable to expect Doug's group could go.

To review, soon after the most powerful invention of his century appeared, Doug Engelbart understood what few others did—not just that computers could enable fantastic especially-mutually-improving tools, but lots of detail about what those tools would look like. Doug correctly saw that computer tools have many synergies, offering tighter than usual loops of self-improvement. He envisioned a rapidly self-improving team focused on developing tools to help them develop better tools, and then actually oversaw a skilled team pursuing his vision for many years. This team created working systems embodying dramatically prescient features, and wowed the computer world with a dramatic demo.

Wasn't this a perfect storm for a tool-takeoff scenario? What odds would have been reasonable to assign to Doug's team "taking over the world"?

* * *

See original post for all comments.

* * *

1. Douglas C. Engelbart, *Augmenting Human Intellect: A Conceptual Framework*, technical report (Menlo Park, CA: Stanford Research Institute, October 1962), http://www.dougengelbart.org/pubs/augment-3906.html.

2. *Wikipedia*, s.v. "Douglas Engelbart," accessed November 12, 2008, http://en.wikipedia.org/w/index.php?title=Douglas_Engelbart&oldid=251218108.

3. *Wikipedia*, s.v. "The Mother of All Demos," accessed October 1, 2008, http://en.wikipedia.org/w/index.php?title=The_Mother_of_All_Demos&oldid=242319216.

3

Friendly Teams

Robin Hanson

15 November 2008

Wednesday I described *UberTool*, an imaginary firm planning to push a set of tools through a rapid mutual-improvement burst until they were in a position to basically "take over the world." I asked when such a plan could be reasonable.

Thursday I noted that Doug Engelbart understood in '62 that computers were the most powerful invention of his century, and could enable especially-mutually-improving tools. He understood lots of detail about what those tools would look like long before others, and oversaw a skilled team focused on his tools-improving-tools plan. That team pioneered graphic user interfaces and networked computers and in '68 introduced the world to the mouse, videoconferencing, email, and the web.

I asked if this wasn't ideal for an *UberTool* scenario, where a small part of an old growth mode "takes over" most of the world via having a head start on a new faster growth mode. Just as humans displaced chimps, farmers displaced hunters, and industry displaced farming, would a group with this much of a head start on such a general better tech have a decent shot at displacing industry folks? And if so, shouldn't the rest of the world have worried about how "friendly" they were?

In fact, while Engelbart's ideas had important legacies, his team didn't come remotely close to displacing much of anything. He lost most of his funding in the early 1970s, and his team dispersed. Even though Engelbart understood key elements of tools that today greatly improve team productivity, his team's tools did not seem to have enabled them to be radically productive, even at the task of improving their tools.

It is not so much that Engelbart missed a few key insights about what computer productivity tools would look like. I doubt it would have made much difference had he traveled in time to see a demo of modern tools. The point is that most tools require lots more than a few key insights to be effective—they also require thousands of small insights that usually accumulate from a large community of tool builders and users.

Small teams have at times suddenly acquired disproportionate power, and I'm sure their associates who anticipated this possibility used the usual human ways to consider that team's "friendliness." But I can't recall a time when such sudden small team power came from an *UberTool* scenario of rapidly mutually improving tools.

Some say we should worry that a small team of AI minds, or even a single mind, will find a way to rapidly improve themselves and take over the world. But what makes that scenario reasonable if the *Uber-Tool* scenario is not?

* * *

Eliezer Yudkowsky

What, in your perspective, distinguishes Doug Engelbart from the two previous occasions in history where a world takeover successfully occurred? I'm not thinking of farming or industry, of course.

Robin Hanson

Eliezer, I discussed what influences transition inequality here.[1] . . .

See original post for all comments.

* * *

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

4

Friendliness Factors

Robin Hanson

16 November 2008

Imagine several firms competing to make the next generation of some product, like a lawn mower or cell phone. What factors influence variance in their product quality (relative to cost)? That is, how much better will the best firm be relative to the average, second best, or worst? Larger variance factors should make competitors worry more that this round of competition will be their last. Here are a few factors:

1. **Resource Variance**—The more competitors vary in resources, the more performance varies.

2. **Cumulative Advantage**—The more prior wins help one win again, the more resources vary.

3. **Grab It First**—If the cost to grab and defend a resource is much less than its value, the first to grab can gain a further advantage.

4. **Competitor Count**—With more competitors, the best exceeds the second best less, but exceeds the average more.

5. **Competitor Effort**—The longer competitors work before their performance is scored, or the more resources they spend, the more scores vary.

6. **Lumpy Design**—The more quality depends on a few crucial choices, relative to many small choices, the more quality varies.

7. **Interdependence**—When firms need inputs from each other, winner gains are also supplier gains, reducing variance.

8. **Info Leaks**—The more info competitors can gain about others' efforts, the more the best will be copied, reducing variance.

9. **Shared Standards**—Competitors sharing more standards and design features in info, process, or product can better understand and use info leaks.

10. **Legal Barriers**—May prevent competitors from sharing standards, info, inputs.

11. **Anti-Trust**—Social coordination may prevent too much winning by a few.

12. **Sharing Deals**—If firms own big shares in each other, or form a co-op, or just share values, they may mind less if others win. Lets them tolerate more variance, but also share more info.

13. **Niche Density**—When each competitor can adapt to a different niche, they may all survive.

14. **Quality Sensitivity**—Demand/success may be very sensitive, or not very sensitive, to quality.

15. **Network Effects**—Users may prefer to use the same product regardless of its quality.

16. [*What factors am I missing? Tell me and I'll extend the list.*]

Some key innovations in history were associated with very high variance in competitor success. For example, our form of life seems to have eliminated all trace of any other forms on Earth. On the other hand, farming and industry innovations were associated with much less variance. I attribute this mainly to info becoming much leakier, in part due to more shared standards, which seems to bode well for our future.

If you worry that one competitor will severely dominate all others in the next really big innovation, forcing you to worry about its "friendliness," you should want to promote factors that reduce success variance. (Though if you cared mainly about the winning performance level, you'd want more variance.)

* * *

Eliezer Yudkowsky

> If you worry that the next really big innovation will be "unfriendly" in the sense of letting one competitor severely dominate all others . . .

This simply isn't the way I use the word "unFriendly." I use it to refer to terminal values and to final behaviors. A single mind that is more powerful than any other on the playing field, but doesn't run around killing people or telling them what to do, can be quite Friendly in both the intuitive sense and the benevolent-terminal-values sense.

Calling this post "Friendliness Factors" rather than "Local vs. Global Takeoff" is needlessly confusing. And I have to seriously wonder—is this the way you had thought I defined "Friendly AI"? If so, this would seem to indicate very little familiarity with my positions at all.

Or are you assuming that a superior tactical position automatically equates to "dominant" behavior in the unpleasant sense, hence "unFriendly" in the intuitive sense? This will be true for many possible goal systems, but not ones that have terminal values that assign low utilities to making people unhappy.

Robin Hanson

Eliezer, yes, sorry—I've just reworded that sentence.

Eliezer Yudkowsky

Okay, with that rewording—i.e., "These are factors that help determine why, how much, what kind of, and how soon you need to worry about Friendliness"—I agree with all factors you have listed. I would add the following:

- **Structure Variance**—the more differently designed competitors are, the more they will vary. Behaves much the same way as Resource Variance and may mitigate against Shared Standards.
- **Recursivity**—the speed at which the "output" of a competitor, in some sense, becomes a resource input or a variant structure.

These factors and the curve of self-optimization implied in Cumulative Advantage are where I put most of my own attention, and it's what I think accounts for human brains taking over but Doug Engelbart failing to do so.

Another factor:

- **Shared Values/Smooth Payoffs**—the more that "competitors" (which are, in this discussion, being described more like runners in a race than business competitors) share each others' values, and the more they are thinking in terms of relatively smooth quantitative payouts and less in terms of being the first to reach the Holy Grail, the more likely they are to share info.

(I.e., this is why Doug Engelbart was more likely to share the mouse with fellow scientists than AI projects with different values are to cooperate.)

Others who think about these topics often put their focus on:

- **Trust-busting**—competitors in aggregate, or a social force outside the set of competitors, try to impose upper limits on power, market share, outlaw certain structures, etc. Has subfactors like Monitoring effectiveness, Enforcement effectiveness and speed, etc.
- **Ambition**—competitors that somehow manage not to want superior positions will probably not achieve them.
- **Compacts**—competitors that can create and keep binding agreements to share the proceeds of risky endeavors will be less unequal afterward.
- **Reproduction**—if successful competitors divide and differentiate they are more likely to create a clade.

Probably not exhaustive, but that's what's coming to mind at the moment.

Eliezer Yudkowsky

- **Rivalness/Exclusivity**—a good design can in principle be used by more than one actor, unless patents prevent it. Versus one AI that takes over all the poorly defended computing power on the Internet may then defend it against other AIs.

Robin Hanson

. . . I edited the list to include many of your suggestions. Not sure I understand "recursivity." I don't see that AIs have more cumulative advantage than human tool teams, and I suspect this CA concept is better broken into components.

See original post for all comments.

5

The Weak Inside View

Eliezer Yudkowsky

18 November 2008

Followup to: The Outside View's Domain

When I met Robin in Oxford for a recent conference, we had a preliminary discussion on the Intelligence Explosion—this is where Robin suggested using production functions. And at one point Robin said something like, "Well, let's see whether your theory's predictions fit previously observed growth-rate curves," which surprised me, because I'd never thought of that at all.

It had never occurred to me that my view of optimization ought to produce quantitative predictions. It seemed like something only an economist would try to do, as 'twere. (In case it's not clear, sentence one is self-deprecating and sentence two is a compliment to Robin—EY)

Looking back, it's not that I made a choice to deal only in qualitative predictions, but that it didn't really occur to me to do it any other way.

Perhaps I'm prejudiced against the Kurzweilian crowd, and their Laws of Accelerating Change and the like. Way back in the distant beginning that feels like a different person, I went around talking about Moore's Law and the extrapolated arrival time of "human-equivalent hardware" à la Moravec. But at some point I figured out that if you weren't exactly reproducing the brain's algorithms, porting cognition to fast serial hardware and to human design instead of evolved adaptation would toss the numbers out the window—and that how much hardware you needed depended on how smart you were—and that sort of thing.

Betrayed, I decided that the whole Moore's Law thing was silly and a corruption of futurism, and I restrained myself to qualitative predictions (and retrodictions) thenceforth.

Though this is to some extent an argument produced after the conclusion, I would explain my reluctance to venture into *quantitative* futurism via the following trichotomy:

- On problems whose pieces are individually *precisely* predictable, you can use the Strong Inside View to calculate a final outcome that has never been seen before—plot the trajectory of the first moon rocket before it is ever launched, or verify a computer chip before it is ever manufactured.

- On problems that are drawn from a barrel of causally similar problems, where human optimism runs rampant and unforeseen troubles are common, the Outside View beats the Inside

View. Trying to visualize the course of history piece by piece will turn out to not (for humans) work so well, and you'll be better off assuming a probable distribution of results similar to previous historical occasions—without trying to adjust for all the reasons why *this* time will be different and better.

- But on problems that are new things under the Sun, where there's a huge change of context and a structural change in underlying causal forces, the Outside View also fails—try to use it, and you'll just get into arguments about what is the proper domain of "similar historical cases" or what conclusions can be drawn therefrom. In this case, the best we can do is use the Weak Inside View—visualizing the causal process—to produce *loose, qualitative conclusions about only those issues where there seems to be lopsided support*.

So to me it seems "obvious" that my view of optimization is only strong enough to produce loose, qualitative conclusions, and that it can only be matched to its retrodiction of history, or wielded to produce future predictions, on the level of qualitative physics.

"Things should speed up here," I could maybe say. But not "The doubling time of this exponential should be cut in half."

I aspire to a deeper understanding of *intelligence* than this, mind you. But I'm not sure that even perfect Bayesian enlightenment would let me predict *quantitatively* how long it will take an AI to solve various problems in advance of it solving them. That might just rest on features of an unexplored solution space which I can't guess in advance, even though I understand the process that searches.

Robin keeps asking me what I'm getting at by talking about some reasoning as "deep" while other reasoning is supposed to be "surface." One thing which makes me worry that something is "surface" is when it involves generalizing a level N feature across a shift in level $N - 1$ causes.

For example, suppose you say, "Moore's Law has held for the last sixty years, so it will hold for the next sixty years, even after the advent of superintelligence" (as Kurzweil seems to believe, since he draws his graphs well past the point where you're buying a billion times human brainpower for $1,000).

Now, if the Law of Accelerating Change were an exogenous, ontologically fundamental, precise physical law, then you wouldn't expect it to change with the advent of superintelligence.

But to the extent that you believe Moore's Law depends on human engineers, and that the timescale of Moore's Law has something to do with the timescale on which human engineers think, then extrapolating Moore's Law across the advent of superintelligence is extrapolating it across a shift in the previous causal generator of Moore's Law.

So I'm worried when I see generalizations extrapolated *across* a change in causal generators not themselves described—i.e., the generalization itself is on the level of the outputs of those generators and doesn't describe the generators directly.

If, on the other hand, you extrapolate Moore's Law out to 2015 because it's been reasonably steady up until 2008—well, Reality is still allowed to say, "So what?" to a greater extent than we can expect to wake up one morning and find Mercury in Mars's orbit. But I wouldn't bet against you, if you just went ahead and drew the graph.

So what's "surface" or "deep" depends on what kind of context shifts you try to extrapolate past.

Robin Hanson said:

> 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.[1] 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.[2]

Why do these transitions occur? Why have they been similar to each other? Are the same causes still operating? Can we expect the next transition to be similar for the same reasons?

One may of course say, "I don't know, I just look at the data, extrapolate the line, and venture this guess—the data is more sure than any hypotheses about causes." And that will be an interesting projection to make, at least.

But you shouldn't be surprised at all if Reality says, "So what?" I mean—real estate prices went up for a long time, and then they went down. And that didn't even require a tremendous shift in the underlying nature and causal mechanisms of real estate.

To stick my neck out further: I am *liable to trust the Weak Inside View over a "surface" extrapolation*, if the Weak Inside View drills down to a deeper causal level and the balance of support is sufficiently lopsided.

I will go ahead and say, "I don't care if you say that Moore's Law has held for the last *hundred* years. Human thought was a primary causal force in producing Moore's Law, and your statistics are all over a domain of human neurons running at the same speed. If you substitute better-designed minds running at a million times human clock speed, the rate of progress ought to speed up—*qualitatively* speaking."

That is, the prediction is without giving precise numbers or supposing that it's still an exponential curve; computation might spike to the limits of physics and then stop forever, etc. But I'll go ahead and say that the rate of technological progress ought to *speed up*, given the said counterfactual intervention on underlying causes to increase the thought speed of engineers by a factor of a million. I'll be downright indignant if Reality says, "So what?" and has the superintelligence make *slower* progress than human engineers instead. It really does seem like an argument so strong that even Reality ought to be persuaded.

It would be interesting to ponder what kind of historical track records have prevailed in such a clash of predictions—trying to extrapolate "surface" features across shifts in underlying causes without speculating about those underlying causes, versus trying to use the Weak Inside View on those causes and arguing that there is "lopsided" support for a qualitative conclusion; in a case where the two came into conflict . . .

. . . kinda hard to think of what that historical case would be, but perhaps I only lack history.

Robin, how surprised would you be if your sequence of long-term exponentials just . . . didn't continue? If the next exponential was too

fast, or too slow, or something other than an exponential? To what degree would you be indignant, if Reality said, "So what?"

* * *

Robin Hanson

It seems reasonable to me to assign a $\sim 1/4$–$1/2$ probability to the previous series not continuing roughly as it has. So it would be only one or two bits of surprise for me.

I suspect it is near time for you to reveal to us your "weak inside view," i.e., the analysis that suggests to you that hand-coded AI is likely to appear in the next few decades, and that it is likely to appear in the form of a single machine suddenly able to take over the world.

See original post for all comments.

* * *

1. Robin Hanson, "Economics of the Singularity," *IEEE Spectrum* 45, no. 6 (2008): 45–50, doi:10.1109/MSPEC.2008.4531461.

2. Hanson, "Outside View of the Singularity."

6

Setting the Stage

Robin Hanson

18 November 2008

As Eliezer and I begin to explore our differing views on intelligence explosion, perhaps I should summarize my current state of mind.

We seem to agree that:

1. Machine intelligence would be a development of almost unprecedented impact and risk, well worth considering now.
2. Feasible approaches include direct hand-coding, based on a few big and lots of little insights, and on emulations of real human brains.
3. Machine intelligence will, more likely than not, appear within a century, even if the progress rate to date does not strongly suggest the next few decades.

4. Many people say silly things here, and we do better to ignore them than to try to believe the opposite.

5. Math and deep insights (especially probability) can be powerful relative to trend fitting and crude analogies.

6. Long-term historical trends are suggestive of future events, but not strongly so.

7. Some should be thinking about how to create "friendly" machine intelligences.

We seem to disagree modestly about the relative chances of the emulation and direct-coding approaches; I think the first and he thinks the second is more likely to succeed first. Our largest disagreement seems to be on the chances that a single hand-coded version will suddenly and without warning change from nearly powerless to overwhelmingly powerful; I'd put it as less than 1% and he seems to put it as over 10%.

At a deeper level, these differences seem to arise from disagreements about what sorts of abstractions we rely on, and on how much we rely on our own personal analysis. My style is more to apply standard methods and insights to unusual topics. So I accept at face value the apparent direct-coding progress to date, and the opinions of most old AI researchers that success there seems many decades off. Since reasonable trend projections suggest emulation will take about two to six decades, I guess emulation will come first.

Though I have physics and philosophy training, and nine years as a computer researcher, I rely most heavily here on abstractions from folks who study economic growth. These abstractions help make

sense of innovation and progress in biology and economies, and can make sense of historical trends, putting apparently dissimilar events into relevantly similar categories. (I'll post more on this soon.) These together suggest a single suddenly superpowerful AI is pretty unlikely.

Eliezer seems to instead rely on abstractions he has worked out for himself, not yet much adopted by a wider community of analysts, nor proven over a history of applications to diverse events. While he may yet convince me to value them as he does, it seems to me that it is up to him to show us how his analysis, using his abstractions, convinces him that, more likely than it might otherwise seem, hand-coded AI will come soon and in the form of a single suddenly superpowerful AI.

* * *

Eliezer Yudkowsky

You give me too much credit. I. J. Good was the one who suggested the notion of an "intelligence explosion" due to the positive feedback of a smart mind making itself even smarter. Numerous other AI researchers believe something similar. I might try to describe the "hard takeoff" concept in a bit more detail but I am hardly its inventor!

Robin Hanson

. . . I didn't mean to imply you had originated the hard takeoff concept. But previous descriptions have been pretty hand-wavy compared to the detail usually worked out when making an argument in the economic growth literature. I want to know what you think is the best presentation and analysis of it, so that I can critique that.

See original post for all comments.

7

The First World Takeover

Eliezer Yudkowsky
19 November 2008

Before Robin and I move on to talking about the Future, it seems to me wise to check if we have disagreements in our view of the Past. Which might be much easier to discuss—and maybe even resolve. So . . .

In the beginning was the Bang. For nine billion years afterward, nothing much happened.

Stars formed and burned for long periods or short periods depending on their structure, but "successful" stars that burned longer or brighter did not pass on their characteristics to other stars. The first replicators were yet to come.

It was the Day of the Stable Things, when your probability of seeing something was given by its probability of accidental formation

times its duration. Stars last a long time; there are many helium atoms.

It was the Era of Accidents, before the dawn of optimization. You'd only expect to see something with forty bits of optimization if you looked through a trillion samples. Something with a thousand bits' worth of functional complexity? You wouldn't expect to find that in the whole universe.

I would guess that, if you were going to be stuck on a desert island and you wanted to stay entertained as long as possible, then you should sooner choose to examine the complexity of the cells and biochemistry of a single Earthly butterfly, over all the stars and astrophysics in the visible universe beyond Earth.

It was the Age of Boredom.

The hallmark of the Age of Boredom was not lack of natural resources—it wasn't that the universe was low on hydrogen—but, rather, the lack of any *cumulative* search. If one star burned longer or brighter, that didn't affect the probability distribution of the next star to form. There was no search but blind search. Everything from scratch, not even looking at the neighbors of previously successful points. Not hill climbing, not mutation and selection, not even discarding patterns already failed. Just a random sample from the same distribution, over and over again.

The Age of Boredom ended with the first replicator.

(Or the first replicator to catch on, if there were failed alternatives lost to history—but this seems unlikely, given the Fermi Paradox; a replicator should be more improbable than that, or the stars would teem with life already.)

Though it might be most dramatic to think of a single RNA strand a few dozen bases long, forming by pure accident after who-knows-how-many chances on who-knows-how-many planets, another class of hypotheses deals with catalytic hypercycles—chemicals whose presence makes it more likely for other chemicals to form, with the arrows happening to finally go around in a circle. If so, RNA would just be a crystallization of that hypercycle into a single chemical that could both take on enzymatic shapes and store information in its sequence for easy replication.

The catalytic hypercycle is worth pondering, since it reminds us that the universe wasn't quite drawing its random patterns from the *same* distribution every time—the formation of a long-lived star made it more likely for a planet to form (if not another star to form), and the formation of a planet made it more likely for amino acids and RNA bases to form in a pool of muck somewhere (if not more likely for planets to form).

In this flow of probability, patterns in one attractor leading to other attractors becoming stronger, there was finally born a *cycle*—perhaps a single strand of RNA, perhaps a crystal in clay, perhaps a catalytic hypercycle—and that was the dawn.

What makes this cycle significant? Is it the large amount of *material* that the catalytic hypercycle or replicating RNA strand could absorb into its pattern?

Well, but any given mountain on Primordial Earth would probably weigh vastly more than the total mass devoted to copies of the first replicator. What effect does mere mass have on optimization?

Suppose the first replicator had a probability of formation of 10^{-30}. If that first replicator managed to make 10,000,000,000 copies

of itself (I don't know if this would be an overestimate or an underestimate for a tidal pool) then this would increase your probability of encountering the replicator pattern by a factor of 10^{10}, the total probability going up to 10^{-20}. (If you were observing "things" at random, that is, and not just on Earth but on all the planets with tidal pools.) So that was a kind of optimization-directed probability flow.

But vastly more important, in the scheme of things, was this—that the first replicator made copies of itself, and some of those copies were errors.

That is, *it explored the neighboring regions of the search space*—some of which contained better replicators—and then those replicators ended up with more probability flowing into them, and explored *their* neighborhoods.

Even in the Age of Boredom there were always regions of attractor space that were the gateways to other regions of attractor space. Stars begot planets, planets begot tidal pools. But that's not the same as a replicator begetting a replicator—it doesn't search a *neighborhood*, find something that better matches a criterion (in this case, the criterion of effective replication), and then search *that* neighborhood, over and over. (x2)

This did require a certain amount of raw material to act as replicator feedstock. But the significant thing was not how much material was recruited into the world of replication; the significant thing was the search, and the material just carried out that search. If, somehow, there'd been some way of doing the same search without all that raw material—if there'd just been a little beeping device that determined how well a pattern *would* replicate, and incremented a binary number representing "how much attention" to pay to that pattern, and then

searched neighboring points in proportion to that number—well, that would have searched just the same. It's not something that evolution *can* do, but if it happened, it would generate the same information.

Human brains routinely outthink the evolution of whole species, species whose net weights of biological material outweigh a human brain a million times over—the gun against a lion's paws. It's not the amount of raw material, it's the search.

In the evolution of replicators, the raw material happens to *carry out* the search—but don't think that the key thing is how much gets produced, how much gets consumed. The raw material is just a way of keeping score. True, even in principle, you do need *some* negentropy and *some* matter to *perform the computation*. But the same search could theoretically be performed with much less material—examining fewer copies of a pattern to draw the same conclusions, using more efficient updating on the evidence. Replicators *happen* to use the number of copies produced of themselves as a way of keeping score.

But what really matters isn't the production, it's the search.

If, after the first primitive replicators had managed to produce a few tons of themselves, you deleted all those tons of biological material, and substituted a few dozen cells here and there from the future—a single algae, a single bacterium—to say nothing of a whole multicellular *C. elegans* roundworm with a 302-neuron *brain*—then Time would leap forward by billions of years, even if the total mass of Life had just apparently shrunk. The *search* would have leapt ahead, and *production* would recover from the apparent "setback" in a handful of easy doublings.

The first replicator was the first great break in History—the first Black Swan that would have been unimaginable by any surface analogy. No extrapolation of previous trends could have spotted it—you'd have had to dive down into causal modeling, in enough detail to visualize the unprecedented search.

Not that I'm saying I *would* have guessed, without benefit of hindsight—if somehow I'd been there as a disembodied and unreflective spirit, knowing only the previous universe as my guide—having no highfalutin concepts of "intelligence" or "natural selection" because those things didn't exist in my environment—and I had no mental mirror in which to see *myself*. And indeed, who *should* have guessed it with short of godlike intelligence? When all the previous history of the universe contained no break in History that sharp? The replicator was the *first* Black Swan.

Maybe I, seeing the first replicator as a disembodied unreflective spirit, would have said, "Wow, what an amazing notion—some of the things I see won't form with high probability, or last for long times—they'll be things that are good at copying themselves, instead. It's the new, third reason for seeing a lot of something!" But would I have been imaginative enough to see the way to amoebas, to birds, to humans? Or would I have just expected it to hit the walls of the tidal pool and stop?

Try telling a disembodied spirit who had watched the whole history of the universe *up to that point* about the birds and the bees, and they would think you were *absolutely and entirely out to lunch*. For nothing *remotely like that* would have been found anywhere else in the universe—and it would obviously take an exponential and *ridiculous* amount of time to accidentally form a pattern like that, no matter how

good it was at replicating itself once formed—and as for it happening many times over in a connected ecology, when the first replicator in the tidal pool took such a long time to happen—why, that would just be *madness*. The Absurdity Heuristic would come into play. Okay, it's neat that a little molecule can replicate itself—but this notion of a "squirrel" is *insanity*. So far beyond a Black Swan that you can't even call it a swan anymore.

That first replicator took over the world—in what sense? Earth's crust, Earth's magma, far outweighs its mass of Life. But Robin and I both suspect, I think, that the fate of the universe, and all those distant stars that outweigh us, will end up shaped by Life. So that the universe ends up hanging quite heavily on the existence of that first replicator, and *not* on the counterfactual states of any particular other molecules nearby . . . In that sense, a small handful of atoms once seized the reins of Destiny.

How? How did the first replicating pattern take over the world? Why didn't all those other molecules get an equal vote in the process?

Well, that initial replicating pattern was doing *some* kind of search—*some* kind of optimization—and nothing else in the Universe was even *trying*. Really it was evolution that took over the world, not the first replicating pattern per se—you don't see many copies of it around anymore. But still, once upon a time the thread of Destiny was seized and concentrated and spun out from a small handful of atoms.

The first replicator did not set in motion a *clever* optimization process. Life didn't even have sex yet, or DNA to store information at very high fidelity. But the rest of the Universe had zip. In the kingdom of blind chance, the myopic optimization process is king.

Issues of "sharing improvements" or "trading improvements" wouldn't even arise—there were no partners from outside. All the agents, all the actors of our modern world, are descended from that first replicator, and none from the mountains and hills.

And that was the story of the First World Takeover, when a shift in the *structure* of optimization—namely, moving from no optimization whatsoever to natural selection—produced a stark discontinuity with previous trends and squeezed the flow of the whole universe's destiny through the needle's eye of a single place and time and pattern.

That's Life.

* * *

Robin Hanson

Eliezer, I can't imagine you really think I disagree with anything important in the above description. I do think it more likely than not that life started before Earth, and so it may have been much less than nine billion years when nothing happened. But that detail hardly matters to the overall picture here.

Eliezer Yudkowsky

Robin, I didn't imagine you would disagree with my history, but I thought you might disagree with my interpretation or emphasis.

Robin Hanson

Eliezer, as someone who has been married for twenty-one years, I know better than to try to pick fights about tone or emphasis when more direct and clear points of disagreement can be found. :)

See original post for all comments.

8

Abstraction, Not Analogy

Robin Hanson

19 November 2008

I'm not that happy with framing our analysis choices here as "surface analogies" versus "inside views." More useful, I think, to see this as a choice of abstractions. An abstraction (Wikipedia) neglects some details to emphasize others. While random abstractions are useless, we have a rich library of useful abstractions tied to specific useful insights.

For example, consider the oldest known tool, the hammer (Wikipedia). To understand how well an ordinary hammer performs its main function, we can abstract from details of shape and materials. To calculate the kinetic energy it delivers, we need only look at its length, head mass, and recoil energy percentage (given by its bending strength). To check that it can be held comfortably, we need the handle's radius, surface coefficient of friction, and shock absorption

ability. To estimate error rates we need only consider its length and head diameter.

For other purposes, we can use other abstractions:

- To see that it is not a good thing to throw at people, we can note it is heavy, hard, and sharp.
- To see that it is not a good thing to hold high in a lightning storm, we can note it is long and conducts electricity.
- To evaluate the cost to carry it around in a tool kit, we consider its volume and mass.
- To judge its suitability as decorative wall art, we consider its texture and color balance.
- To predict who will hold it when, we consider who owns it, and who they know.
- To understand its symbolic meaning in a story, we use a library of common hammer symbolisms.
- To understand its early place in human history, we consider its easy availability and the frequent gains from smashing open shells.
- To predict when it is displaced by powered hammers, we can focus on the cost, human energy required, and weight of the two tools.
- To understand its value and cost in our economy, we can focus on its market price and quantity.

- [*I'm sure we could extend this list.*]

Whether something is "similar" to a hammer depends on whether it has similar *relevant* features. Comparing a hammer to a mask based on their having similar texture and color balance is mere "surface analogy" for the purpose of calculating the cost to carry it around, but is a "deep inside" analysis for the purpose of judging its suitability as wall art. The issue is which abstractions are how useful for which purposes, not which features are "deep" vs. "surface."

Minds are so central to us that we have an enormous range of abstractions for thinking about them. Add that to our abstractions for machines and creation stories, and we have a truly enormous space of abstractions for considering stories about creating machine minds. The issue isn't so much whether any one abstraction is deep or shallow, but whether it is appropriate to the topic at hand.

The future story of the creation of designed minds must of course differ in exact details from everything that has gone before. But that does not mean that nothing before is informative about it. The whole point of abstractions is to let us usefully compare things that are different, so that insights gained about some become insights about the others.

Yes, when you struggle to identify relevant abstractions you may settle for analogizing, i.e., attending to commonly interesting features and guessing based on feature similarity. But not all comparison of different things is analogizing. Analogies are bad not because they use "surface" features, but because the abstractions they use do not offer enough relevant insight for the purpose at hand.

I claim academic studies of innovation and economic growth offer relevant abstractions for understanding the future creation of machine minds, and that in terms of these abstractions the previous major transitions, such as humans, farming, and industry, are relevantly similar. Eliezer prefers "optimization" abstractions. The issue here is evaluating the suitability of these abstractions for our purposes.

* * *

Eliezer Yudkowsky

. . . The dawn of life, considered as a *complete* event, could not have had its properties predicted by similarity to any other *complete* event before it.

But you could, for example, have dropped down to modeling the world on the level of atoms, which would go on behaving similarly to all the other atoms ever observed. It's just that the compound of atoms wouldn't behave similarly to any other compound, with respect to the aspects we're interested in (Life Go FOOM).

You could say, "Probability is flowing between regions of pattern space, the same as before; but look, now there's a cycle; therefore there's this *new* thing going on called *search*." There wouldn't be any *search* in history to analogize to, but there would be (on a lower level of granularity) patterns giving birth to other patterns: stars to planets and the like.

Causal modeling can tell us about things that are not similar *in their important aspect* to any other compound thing in history, provided that they are made out of sufficiently similar *parts* put together in a new structure.

I also note that referring to "humans, farming, and industry" as "the previous major transitions" is precisely the issue at hand—is this an abstraction that's going to give us a good prediction of "self-improving AI" by direct induction/extrapolation, or not?

I wouldn't begin to compare the shift from *non-recursive optimization to recursive optimization* to anything else except the dawn of life—and that's not suggesting that we could do inductive extrapolation, it's just a question of "How

large an event?" There *isn't* anything directly similar to a self-improving AI, in my book; it's a new thing under the Sun, "like replication once was," but not at all the same sort of hammer—if it was, it wouldn't be a new thing under the Sun.

Robin Hanson

Eliezer, have I completely failed to communicate here? You have previously said nothing is similar enough to this new event for analogy to be useful, so all we have is "causal modeling" (though you haven't explained what you mean by this in this context). This post is a reply saying, no, there are more ways using abstractions; analogy and causal modeling are two particular ways to reason via abstractions, but there are many other ways. But here again in the comments you just repeat your previous claim. Can't you see that my long list of ways to reason about hammers isn't well summarized by an analogy vs. causal modeling dichotomy, but is better summarized by noting they use different abstractions? I am of course open to different way to conceive of "the previous major transitions." I have previously tried to conceive of them in terms of sudden growth speedups.

See original post for all comments.

9

Whence Your Abstractions?

Eliezer Yudkowsky

20 November 2008

Reply to: Abstraction, Not Analogy

Robin asks:

> Eliezer, have I completely failed to communicate here? You have previously said nothing is similar enough to this new event for analogy to be useful, so all we have is "causal modeling" (though you haven't explained what you mean by this in this context). This post is a reply saying, no, there are more ways using abstractions; analogy and causal modeling are two particular ways to reason via abstractions, but there are many other ways.

Well . . . it shouldn't be surprising if you've communicated less than you thought. Two people, both of whom know that disagreement is

not allowed, have a persistent disagreement. It doesn't excuse anything, but—wouldn't it be *more* surprising if their disagreement rested on intuitions that were easy to convey in words, and points readily dragged into the light?

I didn't think from the beginning that I was succeeding in communicating. Analogizing Doug Engelbart's mouse to a self-improving AI is for me such a flabbergasting notion—indicating such completely different ways of thinking about the problem—that I am trying to step back and find the differing sources of our differing intuitions.

(Is that such an odd thing to do, if we're really following down the path of not agreeing to disagree?)

"Abstraction," for me, is a word that means a partitioning of possibility—a boundary around possible things, events, patterns. They are in no sense neutral; they act as signposts saying "lump these things together for predictive purposes." To use the word "singularity" as ranging over human brains, farming, industry, and self-improving AI is very nearly to finish your thesis right there.

I wouldn't be surprised to find that, in a real AI, 80% of the actual computing crunch goes into drawing the right boundaries to make the actual reasoning possible. The question "Where do abstractions come from?" cannot be taken for granted.

Boundaries are drawn by appealing to other boundaries. To draw the boundary "human" around things that wear clothes and speak language and have a certain shape, you must have previously noticed the boundaries around clothing and language. And your visual cortex already has a (damned sophisticated) system for categorizing visual scenes into shapes, and the shapes into categories.

It's very much worth distinguishing between boundaries drawn by noticing a set of similarities, and boundaries drawn by reasoning about causal interactions.

There's a big difference between saying, "I predict that Socrates, *like other humans I've observed*, will fall into the class of 'things that die when drinking hemlock' " and saying, "I predict that Socrates, whose biochemistry I've observed to have this-and-such characteristics, will have his neuromuscular junction disrupted by the coniine in the hemlock—even though I've never seen that happen, I've seen lots of organic molecules and I know how they behave."

But above all—ask where the abstraction comes from!

To see a hammer is not good to hold high in a lightning storm, we draw on pre-existing objects that you're not supposed to hold electrically conductive things to high altitudes—this is a predrawn boundary, found by us in books; probably originally learned from experience and then further explained by theory. We just test the hammer to see if it fits in a pre-existing boundary, that is, a boundary we drew before we ever thought about the hammer.

To evaluate the cost to carry a hammer in a tool kit, you probably visualized the process of putting the hammer in the kit, and the process of carrying it. Its mass determines the strain on your arm muscles. Its volume and *shape*—not just "volume," as you can see as soon as that is pointed out—determine the difficulty of fitting it into the kit. You said, "volume and mass," but that was an approximation, and as soon as I say, "volume and mass and shape," you say, "Oh, of course that's what I meant"—based on a causal visualization of trying to fit some weirdly shaped object into a toolkit, or, e.g., a thin ten-foot pin of low volume and high annoyance. So you're redrawing

the boundary based on a causal visualization which shows that other characteristics can be relevant *to the consequence you care about*.

None of your examples talk about drawing *new* conclusions about the hammer by *analogizing it to other things* rather than directly assessing its characteristics in their own right, so it's not all that good an example when it comes to making predictions about self-improving AI by putting it into a group of similar things that includes farming or industry.

But drawing that particular boundary would already rest on *causal* reasoning that tells you which abstraction to use. Very much an Inside View, and a Weak Inside View, even if you try to go with an Outside View after that.

Using an "abstraction" that covers such massively different things will often be met by a differing intuition that makes a different abstraction, *based on a different causal visualization* behind the scenes. That's what you want to drag into the light—not just say, "Well, I expect this Transition to resemble past Transitions."

Robin said:

> I am of course open to different way to conceive of "the previous major transitions." I have previously tried to conceive of them in terms of sudden growth speedups.

Is that the root source for your abstraction—"things that do sudden growth speedups"? I mean . . . is that really what you want to go with here?

* * *

Robin Hanson

Everything is new to us at some point; we are always trying to make sense of new things by using the abstractions we have collected from trying to understand all the old things.

We are always trying to use our best abstractions to directly assess their characteristics in their own right. Even when we use analogies that is the goal. I said the abstractions I rely on most here come from the economic growth literature. They are not just some arbitrary list of prior events.

Robin Hanson

To elaborate, as I understand it a distinctive feature of your scenario is a sudden growth speedup, due to an expanded growth feedback channel. This is the growth of an overall capability of a total mostly autonomous system whose capacity is mainly determined by its "knowledge," broadly understood. The economic growth literature has many useful abstractions for understanding such scenarios. These abstractions have been vetted over decades by thousands of researchers, trying to use them to understand other systems "like" this, at least in terms of these abstractions.

See original post for all comments.

Part II

Main Sequence

10

AI Go Foom

Robin Hanson

10 November 2008

> It seems to me that it is up to [Eliezer] to show us how his analysis, using his abstractions, convinces him that, more likely than it might otherwise seem, hand-coded AI will come soon and in the form of a single suddenly superpowerful AI.

As this didn't prod a response, I guess it is up to me to summarize Eliezer's argument as best I can, so I can then respond. Here goes:

> A machine intelligence can directly rewrite its *entire* source code and redesign its entire physical hardware. While human brains can in principle modify themselves arbitrarily, in practice our limited understanding of ourselves means we mainly only change ourselves by thinking new thoughts. All else equal this means that machine brains have an advantage in improving themselves.

> A mind without arbitrary capacity limits, which focuses on improving itself, can probably do so indefinitely. The growth rate of its "intelligence" may be slow when it is dumb, but gets faster as it gets smarter. This growth rate also depends on how many parts of itself it can usefully change. So all else equal, the growth rate of a machine intelligence must be greater than the growth rate of a human brain.
>
> No matter what its initial disadvantage, a system with a faster growth rate eventually wins. So if the growth-rate advantage is large enough then yes, a single computer could well go in a few days from less than human intelligence to so smart it could take over the world. QED.

So, Eliezer, is this close enough to be worth my response? If not, could you suggest something closer?

* * *

Eliezer Yudkowsky

Well, the format of my thesis is something like:

> When you break down the history of optimization into things like optimization resources, optimization efficiency, and search neighborhood and come up with any reasonable set of curves fit to the observed history of optimization so far, including the very few points where object-level innovations have increased optimization efficiency, and then you try to fit the same curves to an AI that is putting a large part of its present idea-production flow into direct feedback to increase optimization efficiency (unlike human minds or any other process witnessed heretofore), then you get a curve which is either flat (below a certain threshold) or FOOM (above that threshold).

If that doesn't make any sense, it's cuz I was rushed.

Roughly . . . suppose you have a flat linear line, and this is what happens when you have a laborer pushing on a wheelbarrow at constant speed. Now suppose that the wheelbarrow's speed is proportional to the position to which it has been pushed so far. Folding a linear graph in on itself will produce an exponential graph. What we're doing is, roughly, taking the graph of humans being pushed on by evolution, and science being pushed on by humans, and folding that graph in on itself. The justification for viewing things this way has to do with asking questions like "Why did EURISKO run out of steam?" and "Why can't you keep running an optimizing compiler on its own source code to get something faster and faster?" and considering the degree to which meta-level functions can get encapsulated or improved by object-level pressures, which determine the strength of the connections in the positive feedback loop.

I was rushed, so don't blame me if that doesn't make sense either.

Consider that as my justification for trying to answer the question in a post, rather than a comment.

It seems to me that we are viewing this problem from *extremely* different angles, which makes it more obvious to each of us that the other is just plain wrong than that we trust in the other's rationality; and this is the result of the persistent disagreement. It also seems to me that you expect that you know what I will say next, and are wrong about this, whereas I don't feel like I know what you will say next. It's that sort of thing that makes me reluctant to directly jump to your point in opinion space having assumed that you already took mine fully into account.

Robin Hanson

. . . Your story seems to depend crucially on what counts as "object" vs. "meta" (= "optimization efficiency") level innovations. It seems as if you think object ones don't increase growth rates while meta ones do. The economic growth literature pays close attention to which changes increase growth rates and which do not. So I will be paying close attention to how you flesh out your distinction and how it compares with the apparently similar economic growth distinction.

See original post for all comments.

11

Optimization and the Intelligence Explosion

Eliezer Yudkowsky

23 June 2008

Lest anyone get the wrong impression, I'm juggling multiple balls right now and can't give the latest Intelligence Explosion debate as much attention as it deserves. But lest I annoy my esteemed co-blogger, here is a down payment on my views of the Intelligence Explosion—needless to say, all this is coming way out of order in the posting sequence, but here goes . . .

Among the topics I haven't dealt with yet, and will have to introduce here very quickly, is the notion of an optimization process. Roughly, this is the idea that your power as a mind is your ability to hit small targets in a large search space—this can be either the space of possible futures (planning) or the space of possible designs (invention). Suppose you have a car, and suppose we already know that

your preferences involve travel. Now suppose that you take all the parts in the car, or all the atoms, and jumble them up at random. It's very unlikely that you'll end up with a travel artifact at all, even so much as a wheeled cart—let alone a travel artifact that ranks as high in your preferences as the original car. So, relative to your preference ordering, the car is an extremely *improbable* artifact; the power of an optimization process is that it can produce this kind of improbability.

You can view both intelligence and natural selection as special cases of *optimization*: Processes that hit, in a large search space, very small targets defined by implicit preferences. Natural selection prefers more efficient replicators. Human intelligences have more complex preferences. Neither evolution nor humans have consistent utility functions, so viewing them as "optimization processes" is understood to be an approximation. You're trying to get at the *sort of work being done*, not claim that humans or evolution do this work *perfectly*.

This is how I see the story of life and intelligence—as a story of improbably good designs being produced by optimization processes. The "improbability" here is improbability relative to a random selection from the design space, not improbability in an absolute sense—if you have an optimization process around, then "improbably" good designs become probable.

Obviously I'm skipping over a lot of background material here; but you can already see the genesis of a clash of intuitions between myself and Robin. Robin's looking at populations and resource utilization. I'm looking at production of improbable patterns.

Looking over the history of optimization on Earth up until now, the first step is to conceptually separate the meta level from the object

level—separate the *structure of optimization* from *that which is optimized*.

If you consider biology in the absence of hominids, then on the object level we have things like dinosaurs and butterflies and cats. On the meta level we have things like natural selection of asexual populations, and sexual recombination. The object level, you will observe, is rather more complicated than the meta level. Natural selection is not an *easy* subject and it involves math. But if you look at the anatomy of a whole cat, the cat has dynamics immensely more complicated than "mutate, recombine, reproduce."

This is not surprising. Natural selection is an *accidental* optimization process that basically just started happening one day in a tidal pool somewhere. A cat is the *subject* of millions of years and billions of years of evolution.

Cats have brains, of course, which operate to learn over a lifetime; but at the end of the cat's lifetime that information is thrown away, so it does not accumulate. The cumulative effects of cat brains upon the world as optimizers, therefore, are relatively small.

Or consider a bee brain, or a beaver brain. A bee builds hives, and a beaver builds dams; but they didn't figure out how to build them from scratch. A beaver can't figure out how to build a hive; a bee can't figure out how to build a dam.

So animal brains—up until recently—were not major players in the planetary game of optimization; they were *pieces* but not *players*. Compared to evolution, brains lacked both generality of optimization power (they could not produce the amazing range of artifacts produced by evolution) and cumulative optimization power (their

products did not accumulate complexity over time). For more on this theme see "Protein Reinforcement and DNA Consequentialism."[1]

Very recently, certain animal brains have begun to exhibit both generality of optimization power (producing an amazingly wide range of artifacts, in timescales too short for natural selection to play any significant role) and cumulative optimization power (artifacts of increasing complexity, as a result of skills passed on through language and writing).

Natural selection takes hundreds of generations to do anything and millions of years for *de novo* complex designs. Human programmers can design a complex machine with a hundred interdependent elements in a single afternoon. This is not surprising, since natural selection is an *accidental* optimization process that basically just started happening one day, whereas humans are *optimized* optimizers handcrafted by natural selection over millions of years.

The wonder of evolution is not how well it works, but that it works *at all* without being optimized. This is how optimization bootstrapped itself into the universe—starting, as one would expect, from an extremely inefficient accidental optimization process. Which is not the accidental first replicator, mind you, but the accidental first process of natural selection. Distinguish the object level and the meta level!

Since the dawn of optimization in the universe, a certain structural commonality has held across both natural selection and human intelligence . . .

Natural selection *selects on genes*, but, generally speaking, the genes do not turn around and optimize natural selection. The invention of sexual recombination is an exception to this rule, and so is the

invention of cells and DNA. And you can see both the power and the *rarity* of such events by the fact that evolutionary biologists structure entire histories of life on Earth around them.

But if you step back and take a human standpoint—if you think like a programmer—then you can see that natural selection is *still* not all that complicated. We'll try bundling different genes together? We'll try separating information storage from moving machinery? We'll try randomly recombining groups of genes? On an absolute scale, these are the sort of bright ideas that any smart hacker comes up with during the first ten minutes of thinking about system architectures.

Because natural selection started out so inefficient (as a completely accidental process), this tiny handful of meta-level improvements feeding back in from the replicators—nowhere near as complicated as the structure of a cat—structure the evolutionary epochs of life on Earth.

And *after* all that, natural selection is *still* a blind idiot of a god. Gene pools can evolve to extinction, despite all cells and sex.

Now natural selection does feed on itself in the sense that each new adaptation opens up new avenues of further adaptation; but that takes place on the object level. The gene pool feeds on its own complexity—but only thanks to the protected interpreter of natural selection that runs in the background and is not itself rewritten or altered by the evolution of species.

Likewise, human beings invent sciences and technologies, but we have not *yet* begun to rewrite the protected structure of the human brain itself. We have a prefrontal cortex and a temporal cortex and a cerebellum, just like the first inventors of agriculture. We haven't

started to genetically engineer ourselves. On the object level, science feeds on science, and each new discovery paves the way for new discoveries—but all that takes place with a protected interpreter, the human brain, running untouched in the background.

We have meta-level inventions like science that try to instruct humans in how to think. But the first person to invent Bayes's Theorem did not become a Bayesian; they could not rewrite themselves, lacking both that knowledge and that power. Our significant innovations in the art of thinking, like writing and science, are so powerful that they structure the course of human history; but they do not rival the brain itself in complexity, and their effect upon the brain is comparatively shallow.

The present state of the art in rationality training is not sufficient to turn an arbitrarily selected mortal into Albert Einstein, which shows the power of a few minor genetic quirks of brain design compared to all the self-help books ever written in the twentieth century.

Because the brain hums away invisibly in the background, people tend to overlook its contribution and take it for granted, and talk as if the simple instruction to "test ideas by experiment" or the $p < 0.05$ significance rule were the same order of contribution as an entire human brain. Try telling chimpanzees to test their ideas by experiment and see how far you get.

Now . . . some of us *want* to intelligently design an intelligence that would be capable of intelligently redesigning itself, right down to the level of machine code.

The machine code at first, and the laws of physics later, would be a protected level of a sort. But that "protected level" would not contain the *dynamic of optimization*; the protected levels would not structure

the work. The human brain does quite a bit of optimization on its own, and screws up on its own, no matter what you try to tell it in school. But this *fully wraparound recursive optimizer* would have no protected level that was *optimizing*. All the structure of optimization would be subject to optimization itself.

And that is a sea change which breaks with the entire past since the first replicator, because it breaks the idiom of a protected meta level.

The history of Earth up until now has been a history of optimizers spinning their wheels at a constant rate, generating a constant optimization pressure. And creating optimized products, *not* at a constant rate, but at an accelerating rate, because of how object-level innovations open up the pathway to other object-level innovations. But that acceleration is taking place with a protected meta level doing the actual optimizing. Like a search that leaps from island to island in the search space, and good islands tend to be adjacent to even better islands, but the jumper doesn't change its legs. *Occasionally*, a few tiny little changes manage to hit back to the meta level, like sex or science, and then the history of optimization enters a new epoch and everything proceeds faster from there.

Imagine an economy without investment, or a university without language, a technology without tools to make tools. Once in a hundred million years, or once in a few centuries, someone invents a hammer.

That is what optimization has been like on Earth up until now.

When I look at the history of Earth, I don't see a history of optimization *over time*. I see a history of *optimization power* in, and *optimized products* out. Up until now, thanks to the existence of al-

most entirely protected meta levels, it's been possible to split up the history of optimization into epochs, and, within each epoch, graph the cumulative *object-level* optimization *over time*, because the protected level is running in the background and is not itself changing within an epoch.

What happens when you build a fully wraparound, recursively self-improving AI? Then you take the graph of "optimization in, optimized out," and fold the graph in on itself. Metaphorically speaking.

If the AI is weak, it does nothing, because it is not powerful enough to significantly improve itself—like telling a chimpanzee to rewrite its own brain.

If the AI is powerful enough to rewrite itself in a way that increases its ability to make further improvements, and this reaches all the way down to the AI's full understanding of its own source code and its own design as an optimizer . . . then, even if the graph of "optimization power in" and "optimized product out" looks essentially the same, the graph of optimization over time is going to look completely different from Earth's history so far.

People often say something like, "But what if it requires exponentially greater amounts of self-rewriting for only a linear improvement?" To this the obvious answer is, "Natural selection exerted roughly constant optimization power on the hominid line in the course of coughing up humans; and this doesn't seem to have required exponentially more time for each linear increment of improvement."

All of this is still mere analogic reasoning. A full AGI thinking about the nature of optimization and doing its own AI research and rewriting its own source code is not *really* like a graph of Earth's history folded in on itself. It is a different sort of beast. These analo-

gies are *at best* good for qualitative predictions, and even then I have a large amount of other beliefs not yet posted, which are telling me which analogies to make, *et cetera*.

But if you want to know why I might be reluctant to extend the graph of biological and economic growth *over time*, into the future and over the horizon of an AI that thinks at transistor speeds and invents self-replicating molecular nanofactories and *improves its own source code*, then there is my reason: You are drawing the wrong graph, and it should be optimization power in versus optimized product out, not optimized product versus time. Draw *that* graph, and the results—in what I would call common sense for the right values of "common sense"—are entirely compatible with the notion that a self-improving AI, thinking millions of times faster and armed with molecular nanotechnology, would *not* be bound to one-month economic doubling times. Nor bound to cooperation with large societies of equal-level entities with different goal systems, but that's a separate topic.

On the other hand, if the next Big Invention merely impinged *slightly* on the protected level—if, say, a series of intelligence-enhancing drugs, each good for five IQ points, began to be introduced into society—then I can well believe that the economic doubling time would go to something like seven years, because the basic graphs are still in place, and the fundamental structure of optimization has not really changed all that much, and so you are not generalizing way outside the reasonable domain.

I *really* have a problem with saying, "Well, I don't know if the next innovation is going to be a recursively self-improving AI superintelligence or a series of neuropharmaceuticals, but *whichever one is the*

actual case, I predict it will correspond to an economic doubling time of one month." This seems like sheer Kurzweilian thinking to me, as if graphs of Moore's Law are the fundamental reality and all else a mere shadow. One of these estimates is way too slow and one of them is way too fast—he said, eyeballing his mental graph of "optimization power in vs. optimized product out." If we are going to draw graphs at all, I see no reason to privilege graphs against *times*.

I am juggling many balls right now, and am not able to prosecute this dispute properly. Not to mention that I would prefer to have this whole conversation at a time when I had previously done more posts about, oh, say, the notion of an "optimization process" . . . But let it at least not be said that I am dismissing ideas out of hand without justification, as though I thought them unworthy of engagement; for this I do not think, and I have my own complex views standing behind my Intelligence Explosion beliefs, as one might well expect.

Off to pack, I've got a plane trip tomorrow.

* * *

See original post for all comments.

* * *

1. Eliezer Yudkowsky, "Protein Reinforcement and DNA Consequentialism," *Less Wrong* (blog), November 13, 2007, http://lesswrong.com/lw/l2/protein_reinforcement_and_dna_consequentialism/.

12

Eliezer's Meta-level Determinism

Robin Hanson

23 June 2008

Thank you, esteemed co-blogger Eliezer, for your down payment on future engagement of our clash of intuitions. I too am about to travel and must return to other distractions which I have neglected.

Some preliminary comments. First, to be clear, my estimate of future growth rates based on past trends is intended to be unconditional—I do not claim future rates are independent of which is the next big meta innovation, though I am rather uncertain about which next innovations would have which rates.

Second, my claim to estimate the impact of the next big innovation and Eliezer's claim to estimate a much larger impact from "full AGI" are not yet obviously in conflict—to my knowledge, neither Eliezer nor I claim full AGI will be the next big innovation, nor does

Eliezer argue for a full AGI time estimate that conflicts with my estimated timing of the next big innovation.

Third, it seems the basis for Eliezer's claim that my analysis is untrustworthy "surface analogies" vs. his reliable "deep causes" is that, while I use long-vetted general social science understandings of factors influencing innovation, he uses his own new untested meta-level determinism theory. So it seems he could accept that those not yet willing to accept his new theory might instead reasonably rely on my analysis.

Fourth, while Eliezer outlines his new theory and its implications for overall growth rates, he has as yet said nothing about what his theory implies for transition inequality, and how those implications might differ from my estimates.

OK, now for the meat. My story of everything was told (at least for recent eras) in terms of realized capability, i.e., population and resource use, and was largely agnostic about the specific innovations underlying the key changes. Eliezer's story is that key changes are largely driven by structural changes in optimization processes and their protected meta-levels:

> The history of Earth up until now has been a history of optimizers . . . generating a constant optimization pressure. And creating optimized products, not at a constant rate, but at an accelerating rate, because of how object-level innovations open up the pathway to other object-level innovations. . . . *Occasionally*, a few tiny little changes manage to hit back to the meta level, like sex or science, and then the history of optimization enters a new epoch and everything proceeds faster from there. . . .

> Natural selection selects on genes, but, generally speaking, the genes do not turn around and optimize natural selection. The invention of sexual recombination is an exception to this rule, and so is the invention of cells and DNA. . . . This tiny handful of meta-level improvements feeding back in from the replicators . . . structure the evolutionary epochs of life on Earth. . . .
>
> *Very recently*, certain animal brains have begun to exhibit both generality of optimization power . . . and cumulative optimization power . . . as a result of skills passed on through language and writing. . . . We have meta-level inventions like science that try to instruct humans in how to think. . . . Our significant innovations in the art of thinking, like writing and science, are so powerful that they structure the course of human history; but they do not rival the brain itself in complexity, and their effect upon the brain is comparatively shallow. . . .
>
> Now . . . some of us *want* to intelligently design an intelligence that would be capable of intelligently redesigning itself, right down to the level of machine code. . . . [That] breaks the idiom of a protected meta level. . . . Then even if the graph of "optimization power in" and "optimized product out" looks essentially the same, the graph of optimization over time is going to look completely different from Earth's history so far.

OK, so Eliezer's "meta is max" view seems to be a meta-level determinism view, i.e., that capability growth rates are largely determined, in order of decreasing importance, by innovations at three distinct levels:

1. The dominant optimization process, natural selection, flesh brains with culture, or full AGI

2. Improvements behind the protected meta level of such a process, i.e., cells, sex, writing, science

3. Key "object-level" innovations that open the path for other such innovations

Eliezer offers no theoretical argument for us to evaluate supporting this ranking. But his view does seem to make testable predictions about history. It suggests the introduction of natural selection and of human culture coincided with the very largest capability growth rate increases. It suggests that the next largest increases were much smaller and coincided in biology with the introduction of cells and sex, and in humans with the introduction of writing and science. And it suggests other rate increases were substantially smaller.

The main dramatic events in the traditional fossil record are, according to one source, Any Cells, Filamentous Prokaryotes, Unicellular Eukaryotes, Sexual Eukaryotes, and Metazoans, at 3.8, 3.5, 1.8, 1.1, and 0.6 billion years ago, respectively.[1] Perhaps two of these five events are at Eliezer's level two, and none at level one. Relative to these events, the first introduction of human culture isn't remotely as noticeable. While the poor fossil record means we shouldn't expect a strong correspondence between the biggest innovations and dramatic fossil events, we can at least say this data doesn't strongly support Eliezer's ranking.

Our more recent data is better, allowing clearer tests. The last three strong transitions were humans, farming, and industry, and in terms of growth rate changes these seem to be of similar magnitude. Eliezer seems to predict we will discover the first of these was much stronger than the other two. And while the key causes of these tran-

sitions have long been hotly disputed, with many theories in play, Eliezer seems to pick specific winners for these disputes: intergenerational culture, writing, and scientific thinking.

I don't know enough about the first humans to comment, but I know enough about farming and industry to say Eliezer seems wrong there. Yes, the introduction of writing did roughly correspond in time with farming, but it just doesn't seem plausible that writing caused farming, rather than vice versa. Few could write and what they wrote didn't help farming much. Farming seems more plausibly to have resulted from a scale effect in the accumulation of innovations in abilities to manage plants and animals—we finally knew enough to be able to live off the plants near one place, instead of having to constantly wander to new places.

Also for industry, the key innovation does not seem to have been a scientific way of thinking—that popped up periodically in many times and places, and by itself wasn't particularly useful. My guess is that the key was the formation of networks of science-like specialists, which wasn't possible until the previous economy had reached a critical scale and density.

No doubt innovations can be classified according to Eliezer's scheme, and yes, all else equal, relatively meta innovations are probably stronger; but if as the data above suggests this correlation is much weaker than Eliezer expects, that has important implications for how "full AGI" would play out. Merely having the full ability to change its own meta level need not give such systems anything like the wisdom to usefully make such changes, and so an innovation producing that mere ability might not be among the most dramatic transitions.

* * *

Eliezer Yudkowsky

I feel that I am being perhaps a bit overinterpreted here.

For one thing, the thought of "farming" didn't cross my mind when I was thinking of major innovations, which tells you something about the optimization viewpoint versus the economic viewpoint.

But if I were to try to interpret how farming looks from my viewpoint, it would go like this:

1. Evolution gives humans language, general causal modeling, and long-range planning.

2. Humans figure out that sowing seeds causes plants to grow, realize that this could be helpful six months later, and tell their friends and children. No direct significance to optimization.

3. Some areas go from well-nourished hunter-gatherers to a hundred times as many nutritively deprived farmers. Significance to optimization: there are many more humans around, optimizing . . . maybe slightly worse than they did before, due to poor nutrition. However, you can, in some cases, pour more resources in and get more optimization out, so the object-level trick of farming may have hit back to the meta level in that sense.

4. Farming skills get good enough that people have excess crops, which are stolen by tax collectors, resulting in the creation of governments, cities, and, above all, *professional specialization*.

5. People in cities invent writing.

So that's how I would see the object/meta interplay.

Robin Hanson

Eliezer, so even though you said,

> *Occasionally*, a few tiny little changes manage to hit back to the meta level, like sex or science, and then the history of optimization enters a new epoch and everything proceeds faster from there.

you did not intend at all to say that when we look at the actual times when "everything sped up" we would tend to find such events to have been fundamentally caused by such meta-level changes? Even though you say these "meta-level improvements . . . structure the evolutionary epochs of life on Earth," you did not mean the epochs as observed historically or as defined by when "everything proceeds faster from there"? If there is no relation in the past between speedup causes and these key meta-level changes, why worry that a future meta-level change will cause a speedup then?

See original post for all comments.

* * *

1. Robin Hanson, "Must Early Life Be Easy? The Rhythm of Major Evolutionary Transitions" (Unpublished manuscript, September 23, 1998), accessed August 12, 2012, http://hanson.gmu.edu/hardstep.pdf; J. William Schopf, "Disparate Rates, Differing Fates: Tempo and Mode of Evolution Changed from the Precambrian to the Phanerozoic," *Proceedings of the National Academy of Sciences of the United States of America* 91, no. 15 (1994): 6735–6742, doi:10.1073/pnas.91.15.6735.

13

Observing Optimization

Eliezer Yudkowsky

21 November 2008

Followup to: Optimization and the Intelligence Explosion

In "Optimization and the Intelligence Explosion" I pointed out that history since the first replicator, including human history to date, has *mostly* been a case of *nonrecursive* optimization—where you've got one thingy doing the optimizing, and another thingy getting optimized. When evolution builds a better amoeba, that doesn't change the *structure of evolution*—the mutate-reproduce-select cycle.

But there are exceptions to this rule, such as the invention of sex, which affected the structure of natural selection itself—transforming it to mutate-recombine-mate-reproduce-select.

I was surprised when Robin, in "Eliezer's Meta-Level Determinism" took that idea and ran with it and said:

> His view does seem to make testable predictions about history. It suggests the introduction of natural selection and of human culture coincided with the very largest capability growth rate increases. It suggests that the next largest increases were much smaller and coincided in biology with the introduction of cells and sex, and in humans with the introduction of writing and science. And it suggests other rate increases were substantially smaller.

It hadn't occurred to me to try to derive that kind of testable prediction. Why? Well, partially because I'm not an economist. (Don't get me wrong, it was a virtuous step to try.) But also because the whole issue looked to me like it was a lot more complicated than that, so it hadn't occurred to me to try to directly extract predictions.

What is this "capability growth rate" of which you speak, Robin? There are old, old controversies in evolutionary biology involved here.

Just to start by pointing out the obvious—if there are fixed resources available, only so much grass to be eaten or so many rabbits to consume, then any evolutionary "progress" that we would recognize as producing a better-designed organism may just result in the displacement of the old allele by the new allele—*not* any increase in the population as a whole. It's quite possible to have a new wolf that expends 10% more energy per day to be 20% better at hunting, and in this case the sustainable wolf population will decrease as new wolves replace old.

If I was going to talk about the effect that a meta-level change might have on the "optimization velocity" of natural selection, I would talk about the time for a new adaptation to replace an old adap-

tation after a shift in selection pressures—not the total population or total biomass or total morphological complexity (see below).

Likewise in human history—farming was an important innovation for purposes of optimization, not because it changed the human brain all that much, but because it meant that there were a hundred times as many brains around; and even more importantly, that there were surpluses that could support specialized professions. But many innovations in human history may have consisted of new, improved, more harmful weapons—which would, if anything, have decreased the sustainable population size (though "no effect" is more likely—fewer people means more food means more people).

Or similarly—there's a talk somewhere where either Warren Buffett or Charles Munger mentions how they hate to hear about technological improvements in certain industries—because even if investing a few million can cut the cost of production by 30% or whatever, the barriers to competition are so low that the consumer captures all the gain. So they *have* to invest to keep up with competitors, and the investor doesn't get much return.

I'm trying to measure the optimization velocity of information, not production or growth rates. At the tail end of a very long process, knowledge finally does translate into power—guns or nanotechnology or whatever. But along that long way, if you're measuring the number of material copies of the same stuff (how many wolves, how many people, how much grain), you may not be getting much of a glimpse at optimization velocity. Too many complications along the causal chain.

And this is not just my problem.

Back in the bad old days of pre-1960s evolutionary biology, it was widely taken for granted that there was such a thing as progress, that it proceeded forward over time, and that modern human beings were at the apex.

George Williams's *Adaptation and Natural Selection*, marking the so-called "Williams Revolution" in ev-bio that flushed out a lot of the romanticism and anthropomorphism, spent most of one chapter questioning the seemingly common-sensical metrics of "progress."

Biologists sometimes spoke of "morphological complexity" increasing over time. But how do you measure that, exactly? And at what point in life do you measure it if the organism goes through multiple stages? Is an amphibian more advanced than a mammal, since its genome has to store the information for multiple stages of life?

"There are life cycles enormously more complex than that of a frog," Williams wrote.[1] "The lowly and 'simple' liver fluke" goes through stages that include a waterborne stage that swims using cilia, finds and burrows into a snail, and then transforms into a sporocyst; that reproduces by budding to produce redia; these migrate in the snail and reproduce asexually, then transform into cercaria, which, by wiggling a tail, burrow out of the snail and swim to a blade of grass; there they transform into dormant metacercaria; these are eaten by sheep and then hatch into young flukes inside the sheep, then transform into adult flukes, which spawn fluke zygotes . . . So how "advanced" is that?

Williams also pointed out that there would be a limit to how much information evolution could maintain in the genome against degenerative pressures—which seems like a good principle in practice, though I made some mistakes on *LW* in trying to describe the

theory.[2] Taxonomists often take a current form and call the historical trend toward it "progress," but is that *upward* motion, or just substitution of some adaptations for other adaptations in response to changing selection pressures?

"Today the fishery biologists greatly fear such archaic fishes as the bowfin, garpikes, and lamprey, because they are such outstandingly effective competitors," Williams noted.[3]

So if I were talking about the effect of, e.g., sex as a meta-level innovation, then I would expect, e.g., an increase in the total biochemical and morphological complexity that could be maintained—the lifting of a previous upper bound, followed by an accretion of information. And I might expect a change in the velocity of new adaptations replacing old adaptations.

But to get from there to something that shows up in the fossil record—that's not a trivial step.

I recall reading, somewhere or other, about an ev-bio controversy that ensued when one party spoke of the "sudden burst of creativity" represented by the Cambrian explosion, and wondered why evolution was proceeding so much more slowly nowadays. And another party responded that the Cambrian differentiation was mainly visible *post hoc*—that the groups of animals we have *now* first differentiated from one another *then*, but that *at the time* the differences were not as large as they loom nowadays. That is, the actual velocity of adaptational change wasn't remarkable by comparison to modern times, and only hindsight causes us to see those changes as "staking out" the ancestry of the major animal groups.

I'd be surprised to learn that sex had no effect on the velocity of evolution. It looks like it should increase the speed and number of

substituted adaptations, and also increase the complexity bound on the total genetic information that can be maintained against mutation. But to go from there to just looking at the fossil record and seeing *faster progress*—it's not just me who thinks that this jump to phenomenology is tentative, difficult, and controversial.

Should you expect more speciation after the invention of sex, or less? The first impulse is to say "more," because sex seems like it should increase the optimization velocity and speed up time. But sex also creates mutually reproducing *populations* that share genes among themselves, as opposed to asexual lineages—so might that act as a centripetal force?

I don't even propose to answer this question, just point out that it is actually quite *standard* for the phenomenology of evolutionary theories—the question of which observables are predicted—to be a major difficulty. Unless you're dealing with really *easy* qualitative questions like "Should I find rabbit fossils in the Pre-Cambrian?" (I try to only make predictions about AI, using my theory of optimization, when it looks like an *easy* question.)

Yes, it's more convenient for scientists when theories make easily testable, readily observable predictions. But when I look back at the history of life, and the history of humanity, my first priority is to ask, "What's going on here?" and only afterward see if I can manage to make non-obvious retrodictions. I can't just start with the goal of having a convenient phenomenology. Or similarly: the theories I use to organize my understanding of the history of optimization to date have lots of parameters, e.g., the optimization-efficiency curve that describes optimization output as a function of resource input, or the question of how many low-hanging fruits exist in the neighborhood

of a given search point. Does a larger population of wolves increase the velocity of natural selection, by covering more of the search neighborhood for possible mutations? If so, is that a logarithmic increase with population size, or what?—But I can't just wish my theories into being simpler.

If Robin has a *simpler* causal model, with fewer parameters, that stands directly behind observables and easily coughs up testable predictions, which fits the data well and obviates the need for my own abstractions like "optimization efficiency"—

—then I may have to discard my own attempts at theorizing. But observing a series of material growth modes doesn't contradict a causal model of optimization behind the scenes, because it's a pure phenomenology, not itself a causal model—it doesn't say whether a given innovation had any effect on the optimization velocity of the process that produced future object-level innovations that actually changed growth modes, *et cetera*.

* * *

Robin Hanson

If you can't usefully connect your abstractions to the historical record, I sure hope you have *some* data you can connect them to. Otherwise I can't imagine how you could have much confidence in them.

Eliezer Yudkowsky

Depends on how much stress I want to put on them, doesn't it? If I want to predict that the next growth curve will be an exponential and put bounds

around its doubling time, I need a much finer fit to the data than if I only want to ask obvious questions like "Should I find rabbit fossils in the Pre-Cambrian?" or "Do the optimization curves fall into the narrow range that would permit a smooth soft takeoff?"

Robin Hanson

Eliezer, it seems to me that we can't really debate much more until you actually directly make your key argument. If, at it seems to me, you are still in the process of laying out your views tutorial-style, then let's pause until you feel ready.

Eliezer Yudkowsky

I think we ran into this same clash of styles last time (i.e., back at Oxford). I try to go through things systematically, locate any possible points of disagreement, resolve them, and continue. You seem to want to jump directly to the disagreement and then work backward to find the differing premises. I worry that this puts things in a more disagreeable state of mind, as it were—conducive to feed-backward reasoning (rationalization) instead of feed-forward reasoning.

It's probably also worth bearing in mind that these kinds of metadiscussions are important, since this is something of a trailblazing case here. And that if we really want to set up conditions where we can't agree to disagree, that might imply setting up things in a different fashion than the usual Internet debates.

Robin Hanson

When I attend a talk, I don't immediately jump on anything a speaker says that sounds questionable. I wait until they actually make a main point of their

talk, and then I only jump on points that seem to matter for that main point. Since most things people say actually don't matter for their main point, I find this to be a very useful strategy. I will be very surprised indeed if everything you've said mattered regarding our main point of disagreement.

See original post for all comments.

* * *

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

2. Eliezer Yudkowsky, "Natural Selection's Speed Limit and Complexity Bound," *Less Wrong* (blog), November 4, 2007, http://lesswrong.com/lw/ku/natural_selections_speed_limit_and_complexity/.

3. Williams, *Adaptation and Natural Selection*.

14

Life's Story Continues

Eliezer Yudkowsky

21 November 2008

Followup to: The First World Takeover

As last we looked at the planet, Life's long search in organism space had only just gotten started.

When I try to structure my understanding of the unfolding process of Life, it seems to me that, to understand the *optimization velocity* at any given point, I want to break down that velocity using the following abstractions:

- The searchability of the neighborhood of the current location, and the availability of good/better alternatives in that rough region. Maybe call this the *optimization slope*. Are the fruit low-hanging or high-hanging, and how large are the fruit?

- The *optimization resources*, like the amount of computing power available to a fixed program, or the number of individuals in a population pool.
- The *optimization efficiency*, a curve that gives the amount of search power generated by a given investment of resources, which is presumably a function of the optimizer's structure at that point in time.

Example: If an *object-level* adaptation enables more efficient extraction of resources, and thereby increases the total population that can be supported by fixed available resources, then this increases the *optimization resources* and perhaps the optimization velocity.

How much does optimization velocity increase—how hard does this object-level innovation hit back to the meta level?

If a population is small enough that not all mutations are occurring in each generation, then a larger population decreases the time for a given mutation to show up. If the fitness improvements offered by beneficial mutations follow an exponential distribution, then—I'm not actually doing the math here, just sort of eyeballing—I would expect the optimization velocity to go as log population size, up to a maximum where the search neighborhood is explored thoroughly. (You could test this in the lab, though not just by eyeballing the fossil record.)

This doesn't mean *all* optimization processes would have a momentary velocity that goes as the log of momentary resource investment up to a maximum. Just one mode of evolution would have this character. And even under these assumptions, evolution's *cumulative* optimization wouldn't go as log of *cumulative* resources—the log-pop

curve is just the instantaneous velocity. If we assume that the variance of the neighborhood remains the same over the course of exploration (good points have better neighbors with same variance *ad infinitum*), and that the population size remains the same, then we should see linearly cumulative optimization over time. At least until we start to hit the information bound on maintainable genetic information . . .

These are the sorts of abstractions that I think are required to describe the history of life on Earth in terms of optimization. And I also think that if you don't talk optimization, then you won't be able to understand the causality—there'll just be these mysterious unexplained progress modes that change now and then. In the same way you have to talk natural selection to understand observed evolution, you have to talk optimization velocity to understand observed evolutionary speeds.

The first thing to realize is that meta-level changes are rare, so most of what we see in the historical record will be structured by the *search neighborhoods*—the way that one innovation opens up the way for additional innovations. That's going to be most of the story, not because meta-level innovations are unimportant, but because they are rare.

In "Eliezer's Meta-Level Determinism," Robin lists the following dramatic events traditionally noticed in the fossil record:

> Any Cells, Filamentous Prokaryotes, Unicellular Eukaryotes, Sexual Eukaryotes, Metazoans . . .

And he describes "the last three strong transitions" as:

> Humans, farming, and industry . . .

So let me describe what I see when I look at these events, plus some others, through the lens of my abstractions:

Cells: Force a set of genes, RNA strands, or catalytic chemicals to share a common reproductive fate. (This is the real point of the cell boundary, not "protection from the environment"—it keeps the fruits of chemical labor inside a spatial boundary.) But, as we've defined our abstractions, this is mostly a matter of optimization slope—the quality of the search neighborhood. The advent of cells opens up a tremendously rich new neighborhood defined by *specialization* and division of labor. It also increases the slope by ensuring that chemicals get to keep the fruits of their own labor in a spatial boundary, so that fitness advantages increase. But does it hit back to the meta level? How you define that seems to me like a matter of taste. Cells don't quite change the mutate-reproduce-select cycle. But if we're going to define sexual recombination as a meta-level innovation, then we should also define cellular isolation as a meta-level innovation.

It's worth noting that modern genetic algorithms have not, to my knowledge, reached anything like the level of intertwined complexity that characterizes modern unicellular organisms. Modern genetic algorithms seem more like they're producing individual chemicals, rather than being able to handle individually complex modules. So the cellular transition may be a hard one.

DNA: I haven't yet looked up the standard theory on this, but I would sorta expect it to come *after* cells, since a ribosome seems like the sort of thing you'd have to keep around in a defined spatial location. DNA again opens up a huge new search neighborhood by separating the functionality of chemical shape from the demands of reproducing the pattern. Maybe we should rule that anything which re-

structures the search neighborhood this drastically should count as a hit back to the meta level. (Whee, our abstractions are already breaking down.) Also, DNA directly hits back to the meta level by carrying information at higher fidelity, which increases the total storable information.

Filamentous prokaryotes, unicellular eukaryotes: Meh, so what.

Sex: The archetypal example of a rare meta-level innovation. Evolutionary biologists still puzzle over how exactly this one managed to happen.

Metazoans: The key here is not cells aggregating into colonies with similar genetic heritages; the key here is the controlled specialization of cells with an identical genetic heritage. This opens up a huge new region of the search space, but does not particularly change the nature of evolutionary optimization.

Note that opening a sufficiently huge gate in the search neighborhood may *result* in a meta-level innovation being uncovered shortly thereafter. E.g., if cells make ribosomes possible. One of the main lessons in this whole history is that *one thing leads to another*.

Neurons, for example, may have been the key enabling factor in enabling large-motile-animal body plans, because they enabled one side of the organism to talk with the other.

This brings us to the age of brains, which will be the topic of the next post.

But in the meanwhile, I just want to note that my view is nothing as simple as "meta-level determinism" or "the impact of something is proportional to how meta it is; nonmeta things must have small impacts." Nothing much *meta* happened between the age of sexual meta-

zoans and the age of humans—brains were getting more sophisticated over that period, but that didn't change the nature of evolution.

Some object-level innovations are small, some are medium-sized, some are huge. It's no wonder if you look at the historical record and see a Big Innovation that doesn't look the least bit meta but had a huge impact by itself *and* led to lots of other innovations by opening up a new neighborhood picture of search space. This is allowed. Why wouldn't it be?

You can even get exponential acceleration without anything meta—if, for example, the more knowledge you have, or the more genes you have, the more opportunities you have to make good improvements to them. Without any increase in optimization pressure, the neighborhood gets higher-sloped as you climb it.

My thesis is more along the lines of, "If this is the picture *without* recursion, just imagine what's going to happen when we *add* recursion."

To anticipate one possible objection: I don't expect Robin to disagree that modern civilizations underinvest in meta-level improvements because they take time to yield cumulative effects, are new things that don't have certain payoffs, and, worst of all, tend to be public goods. That's why we don't have billions of dollars flowing into prediction markets, for example. I, Robin, or Michael Vassar could probably think for five minutes and name five major probable-big-win meta-level improvements that society isn't investing in.

So if meta-level improvements are rare in the fossil record, it's not necessarily because it would be *hard* to improve on evolution, or because meta-level improving doesn't accomplish much. Rather, evolution doesn't do anything *because* it will have a long-term payoff a

thousand generations later. Any meta-level improvement also has to grant an object-level fitness advantage in, say, the next two generations, or it will go extinct. This is why we can't solve the puzzle of how sex evolved by pointing directly to how it speeds up evolution. "This speeds up evolution" is just not a valid reason for something to evolve.

Any creative evolutionary biologist could probably think for five minutes and come up with five great ways that evolution could have improved on evolution—but which happen to be more complicated than the wheel, which evolution evolved on only three known occasions (Wikipedia)—or don't happen to grant an *immediate* fitness benefit to a handful of implementers.

* * *

Robin Hanson

Let us agree that the "oomph" from some innovation depends on a lot more than whether it is "meta." Meta innovations may well be on average bigger than the average innovation, but there are many other useful abstractions, such as how much new search space is opened up, that also help to predict an innovation's oomph. And there are many ways in which an innovation can make others easier.

See original post for all comments.

15

Emulations Go Foom

Robin Hanson

22 November 2008

Let me consider the AI-foom issue by painting a (looong) picture of the AI scenario I understand best,[1] whole-brain emulations,[2] which I'll call "bots." Here goes.

When investors anticipate that a bot may be feasible soon, they will estimate their chances of creating bots of different levels of quality and cost, as a function of the date, funding, and strategy of their project. A bot more expensive than any (speedup-adjusted) human wage is of little direct value, but exclusive rights to make a bot costing below most human wages would be worth many trillions of dollars.

It may well be socially cost-effective to start a bot-building project with a 1% chance of success when its cost falls to the trillion-dollar level. But not only would successful investors probably only gain a small fraction of this net social value, it is unlikely any investor group

able to direct a trillion could be convinced the project was feasible—there are just too many smart-looking idiots making crazy claims around.

But when the cost to try a 1% project fell below a billion dollars, dozens of groups would no doubt take a shot. Even if they expected the first feasible bots to be very expensive, they might hope to bring that cost down quickly. Even if copycats would likely profit more than they, such an enormous prize would still be very tempting.

The first priority for a bot project would be to create as much emulation fidelity as affordable, to achieve a functioning emulation, i.e., one you could talk to and so on. Few investments today are allowed a decade of red ink, and so most bot projects would fail within a decade, their corpses warning others about what not to try. Eventually, however, a project would succeed in making an emulation that was clearly sane and cooperative.

How close would its closest competitors then be? If there are many very different plausible approaches to emulation, each project may take a different approach, forcing other projects to retool before copying a successful approach. But enormous investment would be attracted to this race once news got out about even a very expensive successful emulation. As I can't imagine that many different emulation approaches, it is hard to see how the lead project could be much more than a year ahead.

Besides hiring assassins or governments to slow down their competition, and preparing to market bots soon, at this point the main task for the lead project would be to make their bot cheaper. They would try multitudes of ways to cut corners on the emulation implementation, checking to see that their bot stayed sane. I expect several

orders of magnitude of efficiency gains to be found easily at first, but that such gains would quickly get hard to find. While a few key insights would allow large gains, most gains would come from many small improvements.

Some project would start selling bots when their bot cost fell substantially below the (speedup-adjusted) wages of a profession with humans available to scan. Even if this risked more leaks, the vast revenue would likely be irresistible. This revenue might help this group pull ahead, but this product would not be accepted in the marketplace overnight. It might take months or years to gain regulatory approval, to see how to sell it right, and then for people to accept bots into their worlds and to reorganize those worlds to accommodate bots.

The first team to achieve high-fidelity emulation may not be the first to sell bots; competition should be fierce and leaks many. Furthermore, the first to achieve marketable costs might not be the first to achieve much lower costs, thereby gaining much larger revenues. Variation in project success would depend on many factors. These depend not only on who followed the right key insights on high fidelity emulation and implementation corner cutting, but also on abilities to find and manage thousands of smaller innovation and production details, and on relations with key suppliers, marketers, distributors, and regulators.

In the absence of a strong world government or a powerful cartel, it is hard to see how the leader could be so far ahead of its nearest competitors as to "take over the world." Sure, the leader might make many trillions more in profits, so enriching shareholders and local residents as to make Bill Gates look like a tribal chief proud of having more feathers in his cap. A leading nation might even go so far

as to dominate the world as much as Britain, the origin of the Industrial Revolution, once did. But the rich and powerful would at least be discouraged from capricious devastation the same way they have always been, by self-interest.

With a thriving bot economy, groups would continue to explore a variety of ways to reduce bot costs and raise bot value. Some would try larger reorganizations of bot minds. Others would try to create supporting infrastructure to allow groups of sped-up bots to work effectively together to achieve sped-up organizations and even cities. Faster bots would be allocated to priority projects, such as attempts to improve bot implementation and bot inputs, such as computer chips. Faster minds riding Moore's Law and the ability to quickly build as many bots as needed should soon speed up the entire world economy, which would soon be dominated by bots and their owners.

I expect this economy to settle into a new faster growth rate, as it did after previous transitions like humans, farming, and industry. Yes, there would be a vast new range of innovations to discover regarding expanding and reorganizing minds, and a richer economy will be increasingly better able to explore this space, but as usual the easy wins will be grabbed first, leaving harder nuts to crack later. And from my AI experience, I expect those nuts to be very hard to crack, though such a enormously wealthy society may well be up to the task. Of course within a few years of more rapid growth we might hit even faster growth modes, or ultimate limits to growth.

Doug Engelbart was right that computer tools can improve computer tools, allowing a burst of productivity by a team focused on tool improvement, and he even correctly saw the broad features of future computer tools. Nevertheless Doug could not translate this into team

success. Inequality in who gained from computers has been less about inequality in understanding key insights about computers, and more about lumpiness in cultures, competing standards, marketing, regulation, etc.

These factors also seem to me the most promising places to look if you want to reduce inequality due to the arrival of bots. While bots will be a much bigger deal than computers were, inducing much larger inequality, I expect the causes of inequalities to be pretty similar. Some teams will no doubt have leads over others, but info about progress should remain leaky enough to limit those leads. The vast leads that life has gained over nonlife, and humans over nonhumans, are mainly due, I think, to the enormous difficulty of leaking innovation info across those boundaries. Leaky farmers and industrialists had far smaller leads.

Added: Since comments focus on slavery, let me quote myself:

> Would robots be slaves? Laws could conceivably ban robots or only allow robots "born" with enough wealth to afford a life of leisure. But without global and draconian enforcement of such laws, the vast wealth that cheap robots offer would quickly induce a sprawling, unruly black market. Realistically, since modest enforcement could maintain only modest restrictions, huge numbers of cheap (and thus poor) robots would probably exist; only their legal status would be in question. Depending on local politics, cheap robots could be "undocumented" illegals, legal slaves of their creators or owners, "free" minds renting their bodies and services and subject to "eviction" for nonpayment, or free minds saddled with debts and subject to "repossession" for nonpayment. The following conclusions do not much depend on which of these cases is more common.[3]

* * *

Carl Shulman

> In the absence of a strong world government or a powerful cartel, it is hard to see how the leader could be so far ahead of its nearest competitors as to "take over the world."

The first competitor uses some smart people with common ideology and relevant expertise as templates for its bots. Then, where previously there were thousands of experts with relevant skills to be hired to improve bot design, there are now millions with initially exactly shared aims. They buy up much of the existing hardware base (in multiple countries), run copies at high speed, and get another order of magnitude of efficiency or so, while developing new skills and digital nootropics. With their vast resources and shared aims they can effectively lobby and cut deals with individuals and governments worldwide, and can easily acquire physical manipulators (including humans wearing cameras, microphones, and remote-controlled bombs for coercions) and cheaply monitor populations.

Copying a bot template is an easy way to build cartels with an utterly unprecedented combination of cohesion and scale.

Carl Shulman

> A leading nation might even go so far as to dominate the world as much as Britain, the origin of the Industrial Revolution, once did.

A leading nation, with territorial control over a large fraction of all world computing hardware, develops brain emulation via a Manhattan Project. Knowing the power of bots, only carefully selected individuals, with high intelligence, relevant expertise, and loyalty, are scanned. The loyalty of the resulting bots is tested exhaustively (copies can be tested to destruction, their digital brains

scanned directly, etc.), and they can be regularly refreshed from old data, and changes carefully tested for effects on motivation.

Server farms are rededicated to host copies of these minds at varying speeds. Many take control of military robots and automated vehicles, while others robustly monitor the human population. The state is now completely secure against human rebellion, and an attack by foreign powers would mean a nuclear war (as it would today). Meanwhile, the bots undertake intensive research to improve themselves. Rapid improvements in efficiency of emulation proceed from workers with a thousandfold or millionfold speedup, with acquisition of knowledge at high speeds followed by subdivision into many instances to apply that knowledge (and regular pruning/replacement of undesired instances). With billions of person-years of highly intelligent labor (but better, because of the ability to spend computational power on both speed and on instances) they set up rapid infrastructure after a period of days and extend their control to the remainder of the planet.

The bots have remained coordinated in values through regular reversion to saved states, and careful testing of the effects of learning and modification on their values (conducted by previous versions) and we now have a global singleton with the values of the national project. That domination is far more extreme than anything ever achieved by Britain or any other historical empire.

Carl Shulman

> . . . are mainly due, I think, to the enormous difficulty of leaking innovation info across those boundaries.

Keeping some technical secrets for at least a few months is quite commonly done, I think it was Tim Tyler who mentioned Google and Renaissance, and militaries have kept many secrets for quite long periods of time when the people involved supported their organizational aim (it was hard to keep Manhattan Project secrets from the Soviet Union because many of the nuclear scientists

supported Communism, but counterintelligence against the Nazis was more successful).

Robin Hanson

. . . I didn't say secrets are never kept, I said human projects leak info lots more than humans did to chimps. If bot projects mainly seek profit, initial humans to scan will be chosen mainly based on their sanity as bots and high-wage abilities. These are unlikely to be pathologically loyal. Ever watch twins fight, or ideologues fragment into factions? Some would no doubt be ideological, but I doubt early bots—copies of them—will be cooperative enough to support strong cartels. And it would take some time to learn to modify human nature substantially. It is possible to imagine how an economically powerful Stalin might run a bot project, and it's not a pretty sight, so let's agree to avoid the return of that prospect.

Carl Shulman

> If bot projects mainly seek profit, initial humans to scan will be chosen mainly based on their sanity as bots and high-wage abilities.

That's a big if. Unleashing "bots"/uploads means setting off the "crack of a future dawn," creating a new supermajority of sapients, driving wages below human subsistence levels, completely upsetting the global military balance of power, and forcing either disenfranchisement of these entities or a handoff of political power in democracies. With rapidly diverging personalities, and bots spread across national borders, it also means scrabbling for power (there is no universal system of property rights), and war will be profitable for many states. Any upset of property rights will screw over those who have not already been uploaded or whose skills are exceeded by those already uploaded, since there will be no economic motivation to keep them alive.

I very much doubt that any U.S. or Chinese President who understood the issues would fail to nationalize a for-profit firm under those circumstances. Even the CEO of an unmolested firm about to unleash bots on the world would think about whether doing so will result in the rapid death of the CEO and the burning of the cosmic commons, and the fact that profits would be much higher if the bots produced were more capable of cartel behavior (e.g., close friends/family of the CEO, with their friendship and shared values tested after uploading).

> It is possible to imagine how an economically powerful Stalin might run a bot project, and it's not a pretty sight, so let's agree to avoid the return of that prospect.

It's also how a bunch of social democrats, or libertarians, or utilitarians, might run a project, knowing that a very likely alternative is the crack of a future dawn and burning the cosmic commons, with a lot of inequality in access to the future, and perhaps worse. Any state with a lead on bot development that can ensure the bot population is made up of nationalists or ideologues (who could monitor each other) could disarm the world's dictatorships, solve collective action problems like the cosmic commons, etc., while releasing the info would hand the chance to conduct the "Stalinist" operation to other states and groups.

> These are unlikely to be pathologically loyal. Ever watch twins fight, or ideologues fragment into factions? Some would no doubt be ideological, but I doubt early bots—copies of them—will be cooperative enough to support strong cartels. And it would take some time to learn to modify human nature substantially.

They will know that the maintenance of their cartel for a time is necessary to avert the apocalyptic competitive scenario, and I mentioned that even without knowledge of how to modify human nature substantially there are ways to prevent value drift. With shared values and high knowledge and intelligence they can use democratic-type decision procedures amongst themselves and enforce those judgments coercively on each other.

Carl Shulman

> And from my AI experience, I expect those nuts to be very hard to crack, though such a enormously wealthy society may well be up to the task.

When does hand-coded AI come into the picture here? Does your AI experience tell you that if you could spend a hundred years studying relevant work in eight sidereal hours, and then split up into a million copies at a thousandfold speedup, you wouldn't be able to build a superhuman initially hand-coded AI in a sidereal month? Likewise for a million von Neumanns (how many people like von Neumann have worked on AI thus far)? A billion? A trillion? A trillion trillion? All this with working brain emulations that can be experimented upon to precisely understand the workings of human minds and inform the hand-coding?

Also, there are a lot of idle mineral and energy resources that could be tapped on Earth and in the solar system, providing quite a number of additional orders of magnitude of computational substrate (raising the returns to improvements in mind efficiency via standard IP economics). A fully automated nanotech manufacturing base expanding through those untapped resources, perhaps with doubling times of significantly less than a week, will enhance growth with an intense positive feedback with tech improvements.

Eliezer Yudkowsky

Carl Shulman has said much of what needed saying.

> Robin: I'm *sure* they will have some short name other than "human." If not "bots," how about "ems"?

Let's go with "ems" (though what was wrong with "uploads"?)

Whole-brain emulations are not part of the AI family, they are part of the modified-human family with the usual advantages and disadvantages thereof, including lots of smart people that seemed nice at first all slowly going insane in the same way, difficulty of modifying the brainware without superhuman in-

telligence, *unavoidable* ethical difficulties, resentment of exploitation and other standard human feelings, *et cetera*.

> They would try multitudes of ways to cut corners on the emulation implementation, checking to see that their bot stayed sane. I expect several orders of magnitude of efficiency gains to be found easily at first, but that such gains would quickly get hard to find.

Leaving aside that you're describing a completely unethical process—as de Blanc notes, prediction is not advocating, but *some* individual humans and governmental entities often at least *try* to avoid doing things that their era says is very wrong, such as killing millions of people—at the very least an economist should *mention* when a putative corporate action involves torture and murder—

—several orders of magnitude of efficiency gains? Without understanding the underlying software in enough detail to write your own *de novo* AI? Suggesting a whole-bird emulation is one thing, suggesting that you can get several orders of magnitude efficiency improvement out of the bird emulation *without understanding how it works* seems like a much, much stronger claim.

As I was initially reading, I was thinking that I was going to reply in terms of ems being nonrecursive—they're just people in silicon instead of carbon, and I for one don't find an extra eight protons all that impressive. It may or may not be *realistic*, but the scenario you describe is not a Singularity in the sense of either a Vingean event horizon or a Goodian intelligence explosion; it's just more of the same but faster.

But any technology powerful enough to milk a thousandfold efficiency improvement out of upload software, without driving those uploads insane, is powerful enough to *upgrade* the uploads. Which brings us to Cameron's observation:

> What the? Are you serious? Are you talking about self replicating machines of $\geq$ human intelligence or Tamagotchi?

I am afraid that my reaction was much the same as Cameron's. The prospect of biological humans sitting on top of a population of ems that are *smarter, much*

faster, and far more numerous than bios *while having all the standard human drives*, and the bios treating the ems as standard economic valuta to be milked and traded around, and the ems sitting still for this for more than a week of bio time—this does not seem historically realistic. . . .

Robin Hanson

All, this post's scenario *assumes* whole-brain emulation without other forms of machine intelligence. We'll need other posts to explore the chances of this vs. other scenarios, and the consequences of other scenarios. This post was to explore the need for friendliness in this scenario.

Note that most objections here are to my social science, and to ethics some try to read into my wording (I wasn't trying to make any ethical claims). No one has complained, for example, that I've misapplied or ignored optimization abstractions.

I remain fascinated by the common phenomenon wherein intuitive social reasoning seems so compelling to most people that they feel very confident of their conclusions and feel little inclination to listen to or defer to professional social scientists. Carl Shulman, for example, finds it obvious it is in the self-interest of "a leading power with an edge in bot technology and some infrastructure . . . to kill everyone else and get sole control over our future light-cone's natural resources." Eliezer seems to say he agrees. I'm sorry, Carl, but your comments on this post sound like crazy paranoid rants, as if you were Dr. Strangelove pushing the button to preserve our precious bodily fluids. Is there any social scientist out there who finds Carl's claims remotely plausible?

Eliezer, I don't find it obviously unethical to experiment with implementation shortcuts on a willing em volunteer (or on yourself). The several orders of magnitude of gains were relative to a likely-to-be excessively high-fidelity initial emulation (the WBE roadmap agrees with me here I think). I did not assume the ems would be slaves, and I explicitly added to the post before your comment to make that clear. If it matters, I prefer free ems who rent or borrow bodies. Finally, is your objection here really going to be that you can't imagine

a world with vast wealth inequality without the poor multitudes immediately exterminating the rich few? Or does this only happen when many poor think faster than many rich? What kind of social science analysis do you base this conclusion on? . . .

Carl Shulman

> Carl Shulman, for example, finds it obvious it is in the self-interest of "a leading power with an edge in bot technology and some infrastructure . . . to kill everyone else and get sole control over our future light-cone's natural resources.

You are misinterpreting that comment. I was directly responding to your claim that self-interest would restrain capricious abuses, as it seems to me that the ordinary self-interested reasons restraining abuse of outgroups, e.g., the opportunity to trade with them or tax them, no longer apply when their labor is worth less than a subsistence wage, and other uses of their constituent atoms would have greater value. There would be little *self-interested* reason for an otherwise abusive group to rein in such mistreatment, even though plenty of altruistic reasons would remain. For most, I would expect them to initially plan simply to disarm other humans and consolidate power, killing only as needed to preempt development of similar capabilities.

> Finally, is your objection here really going to be that you can't imagine a world with vast wealth inequality without the poor multitudes immediately exterminating the rich few? Or does this only happen when many poor think faster than many rich? What kind of social science analysis do you base this conclusion on?

Empirically, most genocides in the last hundred years have involved the expropriation and murder of a disproportionately prosperous minority group. This is actually a common pattern in situations with much less extreme wealth inequality and difference (than in an upload scenario) between ethnic groups in the modern world:

http://www.amazon.com/World-Fire-Exporting-Democracy-Instability/dp/0385503024

Also, Eliezer's point does not require extermination (although a decision simply to engage in egalitarian redistribution, as is common in modern societies, would reduce humans below the subsistence level, and almost all humans would lack the skills to compete in emulation labor markets, even if free uploading was provided), just that if a CEO expects that releasing uploads into the world will shortly upset the economic system in which any monetary profits could be used, the profit motive for doing so will be weak.

James Miller

> I remain fascinated by the common phenomenon wherein intuitive social reasoning seems so compelling to most people that they feel very confident of their conclusions and feel little inclination to listen to or defer to professional social scientists. Carl Shulman, for example, finds it obvious it is in the self-interest of "a leading power with an edge in bot technology and some infrastructure . . . to kill everyone else and get sole control over our future light-cone's natural resources." Eliezer seems to say he agrees. I'm sorry Carl, but your comments on this post sound like crazy paranoid rants, as if you were Dr. Strangelove pushing the button to preserve our precious bodily fluids. Is there any social scientist out there who finds Carl's claims remotely plausible?

Yes.

Ten people are on an island with a limited supply of food. You die when you run out of food. The longer you live the greater your utility. Any one individual might maximize his utility by killing everyone else.

Ten billion people in a universe with a limited supply of usable energy. You die when you run out of usable energy . . .

Or even worse, post-transition offense turns out to be much, much easier than defense. You get to live forever so long as no one kills you. If you care only

about yourself, don't get a huge amount of utility from being in the company of others, then it would be in your interest to kill everyone else.

Carl is only crazy if you assume that a self-interested person would necessarily get a huge amount of utility from living in the company of others. Post-transition this assumption might not be true.

Carl Shulman

James,

> Ten people are on an island with a limited supply of food. You die when you run out of food. The longer you live the greater your utility. Any one individual might maximize his utility by killing everyone else.

Yes, if a secure governing elite, e.g., the top ten thousand Party Members in North Korea (who are willing to kill millions among the Korean population to better secure their safety and security), could decide between an even distribution of future resources among the existing human population vs. only amongst themselves, I would not be surprised if they took a millionfold increase in expected future well-being. A group with initially noble intentions that consolidated global power could plausibly drift to this position with time, and there are many intermediate cases of ruling elites that are nasty but substantially less so than the DPRK's.

> Or even worse, post-transition offense turns out to be much, much easier than defense.

No, this just leads to disarming others and preventing them from gaining comparable technological capabilities.

Robin Hanson

Carl, consider this crazy paranoid rant:

> Don't be fooled, everything we hold dear is at stake! They are completely and totally dedicated to their plan to rule everything, and will annihilate us as soon as they can. They only pretend to be peaceful now to gain temporary advantages. If we forget this and work with them, instead of dedicating ourselves to their annihilation, they will gain the upper hand and all will be lost. Any little advantage we let them have will be used to build even more advantages, so we must never give an inch. Any slight internal conflict on our side will also give them an edge. We must tolerate no internal conflict and must be willing to sacrifice absolutely everything because they are completely unified and dedicated, and if we falter all is lost.

You are essentially proposing that peace is not possible because everyone will assume that others see this as total war, and so fight a total war themselves. Yes, sometimes there are wars, and sometimes very severe wars, but war is rare and increasingly so. Try instead to imagine choices made by folks who think the chance of war was low.

Eliezer Yudkowsky

Robin, are you seriously dismissing the possibility of conflict between bios and ems?

James Miller

Robin,

War is rare today mostly because it's not beneficial. But under different incentive structures humans are very willing to kill to benefit themselves. For example among the Yanomamö (a primitive tribe in Brazil) more than a third of the men die from warfare.

http://en.wikipedia.org/wiki/Yanomami

If the benefits of engaging in warfare significantly increase your "crazy paranoid rant" becomes rather sound advice.

You wrote, "Try instead to imagine choices made by folks who think the chance of war was low." When I imagine this I think of Neville Chamberlain.

Carl Shulman

> You are essentially proposing that peace is not possible because everyone will assume that others see this as total war, and so fight a total war themselves. Yes, sometimes there are wars, and sometimes very severe wars, but war is rare and increasingly so.

I am not proposing that peace is impossible, but that resolving an unstable arms race, with a winner-take-all technology in sight, requires either coordinating measures such as treaties backed by inspection, or trusting in the motives of the leading developer. I would prefer the former. I do not endorse the ludicrous caricature of in-group bias you present and do not think of biological humans as my morally supreme ingroup (or any particular tribe of biological humans, for that matter). If the parable is supposed to indicate that I am agitating for the unity of an ingroup against an ingroup, please make clear which is supposed to be which.

I am proposing that states with no material interests in peace will tend to be less peaceful, that states with the ability to safely disarm all other states will tend to do so, and that states (which devote minimal resources to assisting foreigners and future generations) will tend to allocate unclaimed resources to their citizens or leadership, particularly when those resources can be used to extend life. It is precisely these tendencies that make it worthwhile to make efforts to ensure that the development and application of these technologies is conducted in a transparent and coordinated way, so that arms races and deadly mistakes can be avoided.

Are you essentially proposing that the governments of the world would *knowingly* permit private and uncontrolled development of a technology that will result in permanent global unemployment (at more than a subsistence wage, without subsidy) for biological humans, render biological humans a

weak and tiny minority on this planet, and completely disrupt the current geopolitical order, as well as possibly burning the cosmic commons and/or causing the extinction of biological humans, when it is possible to exert more control over developments? That seems less likely than governments knowingly permitting the construction and possession of nuclear ICBMs by private citizens.

Robin Hanson

Carl, my point is that this tech is not of a type intrinsically more winner-take-all, unstable-arms-like, or geopolitical-order-disrupting than most any tech that displaces competitors via lower costs. This is nothing like nukes, which are only good for war. Yes, the cumulative effects of more new tech can be large, but this is true for most any new tech. Individual firms and nations would adopt this tech for the same reason they adopt other lower-cost tech; because they profit by doing so. Your talk of extinction and "a weak and tiny minority" are only relevant when you imagine wars.

Robin Hanson

James, I agree that it is *possible* for war to be beneficial. The question is whether *in the specific scenario described in this post* we have good reasons to think it would be. . . .

Eliezer Yudkowsky

Any sufficiently slow FOOM is indistinguishable from an investment opportunity.

Robin Hanson

Eliezer, yes, and so the vast majority of fooms may be slow and not require friendliness. So we need positive arguments why any one foom is an exception to this. . . .

See original post for all comments.

* * *

1. Hanson, "Economics of the Singularity."

2. Anders Sandberg and Nick Bostrom, *Whole Brain Emulation: A Roadmap*, Technical Report, 2008-3 (Future of Humanity Institute, University of Oxford, 2008), http://www.fhi.ox.ac.uk/wp-content/uploads/brain-emulation-roadmap-report1.pdf.

3. Hanson, "Economics of the Singularity."

16

Brain Emulation and Hard Takeoff

Carl Shulman

22 November 2008

The construction of a working brain emulation would require, aside from brain-scanning equipment and computer hardware to test and run emulations on, highly intelligent and skilled scientists and engineers to develop and improve the emulation software. How many such researchers? A billion-dollar project might employ thousands, of widely varying quality and expertise, who would acquire additional expertise over the course of a successful project that results in a working prototype. Now, as Robin says:

> They would try multitudes of ways to cut corners on the emulation implementation, checking to see that their bot stayed sane. I expect several orders of magnitude of efficiency gains to be found easily at first, but that such gains

> would quickly get hard to find. While a few key insights would allow large gains, most gains would come from many small improvements.
>
> Some project would start selling bots when their bot cost fell substantially below the (speedup-adjusted) wages of a profession with humans available to scan. Even if this risked more leaks, the vast revenue would likely be irresistible.

To make further improvements they would need skilled workers up to speed on relevant fields and the specific workings of the project's design. But the project above can now run an emulation at a cost substantially less than the wages it can bring in. In other words, it is now cheaper for the project to run an instance of one of its brain emulation engineers than it is to hire outside staff or collaborate with competitors. This is especially so because an emulation can be run at high speeds to catch up on areas it does not know well, faster than humans could be hired and brought up to speed, and then duplicated many times. The limiting resource for further advances is no longer the supply of expert humans, but simply computing hardware on which to run emulations.

In this situation the dynamics of software improvement are interesting. Suppose that we define the following:

- The stock of knowledge, s, is the number of standardized researcher-years that have been expended on improving emulation design.
- The hardware base, h, is the quantity of computing hardware available to the project in generic units.

- The efficiency level, e, is the effective number of emulated researchers that can be run using one generic unit of hardware.

The first derivative of s will be equal to $h \cdot e$, e will be a function of s, and h will be treated as fixed in the short run. In order for growth to proceed with a steady doubling, we will need e to be a very specific function of s, and we will need a different function for each possible value of h. Reduce it much, and the self-improvement will slow to a crawl. Increase h by an order of magnitude over that and you get an immediate explosion of improvement in software, the likely aim of a leader in emulation development.

How will this hardware capacity be obtained? If the project is backed by a national government, it can simply be given a large fraction of the computing capacity of the nation's server farms. Since the cost of running an emulation is less than high-end human wages, this would enable many millions of copies to run at real-time speeds immediately. Since mere thousands of employees (many of lower quality) at the project had been able to make significant progress previously, even with diminishing returns, this massive increase in the effective size, intelligence, and expertise of the workforce (now vastly exceeding the world AI and neuroscience communities in numbers, average IQ, and knowledge) should be able to deliver multiplicative improvements in efficiency and capabilities. That capabilities multiplier will be applied to the project's workforce, now the equivalent of tens or hundreds of millions of Einsteins and von Neumanns, which can then make further improvements.

What if the project is not openly backed by a major state such as Japan, the U.S., or China? If its possession of a low-cost emula-

tion method becomes known, governments will use national security laws to expropriate the technology, and can then implement the plan above. But if, absurdly, the firm could proceed unmolested, then it could likely acquire the needed hardware by selling services. Robin suggests that

> This revenue might help this group pull ahead, but this product would not be accepted in the marketplace overnight. It might take months or years to gain regulatory approval, to see how to sell it right, and then for people to accept bots into their worlds and to reorganize those worlds to accommodate bots.

But there are many domains where sales can be made directly to consumers across national borders, without emulations ever transferring their data to vulnerable locations. For instance, sped-up emulations could create music, computer games, books, and other art of extraordinary quality and sell it online through a website (held by some pre-existing company purchased by the project or the project's backers) with no mention of the source of the IP. Revenues from these sales would pay for the cost of emulation labor, and the residual could be turned to self-improvement, which would slash labor costs. As costs fell, any direct-to-consumer engagement could profitably fund further research, e.g., phone sex lines using VoIP would allow emulations to remotely earn funds with extreme safety from the theft of their software.

Large amounts of computational power could also be obtained by direct dealings with a handful of individuals. A project could secretly investigate, contact, and negotiate with a few dozen of the most plausible billionaires and CEOs with the ability to provide some server

farm time. Contact could be anonymous, with proof of AI success demonstrated using speedups, e.g., producing complex original text on a subject immediately after a request using an emulation with a thousandfold speedup. Such an individual could be promised the Moon, blackmailed, threatened, or convinced of the desirability of the project's aims.

To sum up:

1. When emulations can first perform skilled labor like brain-emulation design at a cost in computational resources less than the labor costs of comparable human workers, mere thousands of humans will still have been making progress at a substantial rate (that's how they get to cost-effective levels of efficiency).

2. Access to a significant chunk of the hardware available at that time will enable the creation of a work force orders of magnitude larger and with much higher mean quality than a human one still making substantial progress.

3. Improvements in emulation software will multiply the efficacy of the emulated research work force, i.e., the return on investments in improved software scales with the hardware base. When the hardware base is small, each software improvement delivers a small increase in the total research power, which may be consumed by diminishing returns and exhaustion of low-hanging fruit; but when the total hardware base is large, positive feedback causes an intelligence explosion.

4. A project, which is likely to be nationalized if obtrusive, could plausibly obtain the hardware required for an intelligence explosion through nationalization or independent action.

* * *

Robin Hanson

This really represents a basic economic confusion. Having a product that you can sell for more than its cost for you to make gives you profits, i.e., wealth. But having wealth does *not* necessarily give you an advantage at finding new ways to get more wealth. So having an advantage at making ems does *not* necessarily give you an advantage at making cheaper ems. Sure, you can invest in research, but so can everyone else who has wealth. You seem to assume here that groups feel compelled to follow a plan of accumulating a war chest of wealth, reinvesting their wealth in gaining more wealth, because they expect to fight a war. And yes, when people expect and plan for wars, well, wars often result. But that hardly means that if some will gain temporary sources of wealth a war will follow.

Eliezer Yudkowsky

Robin, your reply doesn't seem to take into account the notion of *using em researchers to make cheaper ems*. Whoever has the cheapest ems to start with gets the cheapest research done.

Robin Hanson

Eliezer, you need to review the concept of *opportunity cost*. It is past midnight here, and I'm off to bed now.

Eliezer Yudkowsky

G'night. Sorry, don't see the connection even after being told. I'm not saying that the leading em-builders are getting ems from nowhere without paying opportunity costs, I'm saying they get their ems wholesale instead of retail and this advantage snowballs.

Carl Shulman

> This really represents a basic economic confusion.

Robin, you've made a number of comments along these lines, assuming mistakenly that I am not familiar with standard economic results and literatures and attributing claims to the supposed unfamiliarity, when in fact I am very familiar indeed with economics in general and the relevant results in particular.

I am fully familiar with the decline in casualties from violence in recent centuries, the correlations of peace with economic freedom, democracy, prosperity, etc. I understand comparative advantage and the mistake of mercantilism, self-fulfilling prophecies in arms races, etc., etc. I know you highly value social science and think that other thinkers on futurist topics neglect basic economic results and literatures, and I am not doing so. I agree, and am informed on those literatures.

> But having wealth does *not* necessarily give you an advantage at finding new ways to get more wealth.

In this case we are talking about highly intelligent researchers, engineers, and managers. Those will indeed help you to find new ways to get more wealth!

> So having an advantage at making ems does *not* necessarily give you an advantage at making cheaper ems.

The scenario above explicitly refers to the project that first develops cost-effective ems, not ems in general. Having an advantage at making cost-effective ems means that you can convert cash to improvements in em technology more efficiently by renting hardware and running cost-effective ems on it than by hiring, as I explained above.

> Sure, you can invest in research, but so can everyone else who has wealth.

Initially sole knowledge of cost-effective em design means that you get a vastly, vastly higher return on investment on research expenditures than others do.

> You seem to assume here that groups feel compelled to follow a plan of accumulating a war chest of wealth, reinvesting their wealth in gaining more wealth, because they expect to fight a war.

From a pure profit-maximizing point of view (although again, given the consequences you project from em development, it is absurd to expect that firm would knowingly be allowed to remain private by governments), taking some time to pursue improvement while retaining a monopoly on the relevant IP means hugely increasing the value of one's asset. If the technology is sold the sole control of the IP will be lost, since IP rights are not secure, and many markets where the project would have enjoyed monopoly will become highly competitive, tremendously driving down returns from the asset.

Eliezer Yudkowsky

Many, many information companies choose to keep their source code private and sell services or products, rather than selling the source code itself to get immediate wealth.

Robin Hanson

Eliezer, the opportunity cost of any product is the revenue you would get by selling/renting it to others, not your cost of producing it. If there were a big competitive advantage from buying wholesale over retail from yourself, then firms would want to join large cooperatives where they all buy wholesale from each other, to their mutual advantage. But in fact conglomerates typically suffer from inefficient and inflexible internal pricing contracts; without other

big economies of scope conglomerates are usually more efficient if broken into smaller firms.

Robin Hanson

Carl, I can't win a word war of attrition with you, where each response of size X gets a reply of size $N \cdot X$, until the person who wrote the most crows that most of his points never got a response. I challenge you to write a clear concise summary of your key argument and we'll post it here on *OB*, and I'll respond to that.

James Miller

Carl wrote in a comment:

> Initially sole knowledge of cost-effective em design means that you get a vastly, vastly, higher return on investment on research expenditures than others do.

Let's say that firm A has the cost-effective em design whereas firm B has a cost-ineffective em design. Imagine that it will take firm B lots of time and capital to develop a cost-effective em design.

True, give both firm A and firm B a dollar and firm A could use it to generate more revenue than firm B could.

But if firm B is expected to earn a long-term positive economic profit it could raise all the money it wanted on capital markets. There would be no financial constraint on firm B and thus no financial market advantage to firm A even if firm A could always earn greater accounting profits than firm B.

(Economists define profit taking into account opportunity costs. So let's say I can do X or Y but not both. If X would give me $20 and Y $22 then my economic profit from doing Y is $2. In contrast an accountant would say that doing Y gives you a profit of $22. I'm not assuming that Carl doesn't know this.)

Carl Shulman

> But if firm B is expected to earn a long-term positive economic profit it could raise all the money it wanted on capital markets.

Provided that contract enforcement and property rights are secure, so that lenders believe they will be repaid, and can be approached without resulting in government expropriation. The expropriation concern is why my discussion above focuses on ways to acquire hardware/funds without drawing hostile attention. However, I did mention lending, as "promising the Moon," since while a firm using loan funding to conduct an in-house intelligence explosion could promise absurdly high interest rates, if it were successful creditors would no longer be able to enforce a contractual obligation for repayment through the legal system, and would need to rely on the honor of the debtor.

See original post for all comments.

17

Billion Dollar Bots

James Miller

22 November 2008

Robin presented a scenario in which whole-brain emulations, or what he calls *bots*, come into being. Here is another:

Bots are created with hardware and software. The higher the quality of one input the less you need of the other. Hardware, especially with cloud computing, can be quickly allocated from one task to another. So the first bot might run on hardware worth billions of dollars.

The first bot creators would receive tremendous prestige and a guaranteed place in the history books. So once it becomes possible to create a bot many firms and rich individuals will be willing to create one even if doing so would cause them to suffer a large loss.

Imagine that some group has $300 million to spend on hardware and will use the money as soon as $300 million becomes enough to create a bot. The best way to spend this money would not be to buy a

$300 million computer but to rent $300 million of off-peak computing power. If the group needed only a thousand hours of computing power (which it need not buy all at once) to prove that it had created a bot then the group could have, roughly, $3 billion of hardware for the needed thousand hours.

It's likely that the first bot would run very slowly. Perhaps it would take the bot ten real seconds to think as much as a human does in one second.

Under my scenario the first bot would be wildly expensive. But, because of Moore's Law, once the first bot was created everyone would expect that the cost of bots would eventually become low enough so that they would radically remake society.

Consequently, years before bots come to dominate the economy, many people will come to expect that within their lifetime bots will someday come to dominate the economy. Bot expectations will radically change the world.

I suspect that after it becomes obvious that we could eventually create cheap bots world governments will devote trillions to bot Manhattan Projects. The expected benefits of winning the bot race will be so high that it would be in the self-interest of individual governments to not worry too much about bot friendliness.

The U.S. and Chinese militaries might fall into a bot prisoner's dilemma in which both militaries would prefer an outcome in which everyone slowed down bot development to ensure friendliness yet both nations were individually better off (regardless of what the other military did) taking huge chances on friendliness so as to increase the probability of their winning the bot race.

My hope is that the U.S. will have such a tremendous advantage over China that the Chinese don't try to win the race and the U.S. military thinks it can afford to go slow. But given China's relatively high growth rate I doubt humanity will luck into this safe scenario.

* * *

Robin Hanson

Like Eliezer and Carl, you assume people will assume they are in a total war and act accordingly. There need not be a "race" to "win." I shall have to post on this soon I guess.

James Miller

Robin—in your response post please consider asking, "What would John von Neumann do?" He advocated a first-strike attack on the Soviet Union.

See original post for all comments.

18

Surprised by Brains

Eliezer Yudkowsky

23 November 2008

Followup to: Life's Story Continues

Imagine two agents who've *never seen an intelligence*—including, somehow, themselves—but who've seen the rest of the universe up until now, arguing about what these newfangled "humans" with their "language" might be able to do . . .

> Believer: Previously, evolution has taken hundreds of thousands of years to create new complex adaptations with many working parts. I believe that, thanks to brains and language, we may see a *new* era, an era of *intelligent design*. In this era, complex causal systems—with many interdependent parts that collectively serve a definite function—will be created by the cumulative work of many brains building upon each others' efforts.

Skeptic: I see—you think that brains might have something like a 50% speed advantage over natural selection? So it might take a while for brains to catch up, but after another eight billion years, brains will be in the lead. But this planet's Sun will swell up by then, so—

Believer: *Thirty percent*? I was thinking more like *three orders of magnitude*. With thousands of brains working together and building on each others' efforts, whole complex machines will be designed on the timescale of mere millennia—no, *centuries*!

Skeptic: *What*?

Believer: You heard me.

Skeptic: Oh, come on! There's absolutely no empirical evidence for an assertion like that! Animal brains have been around for hundreds of millions of years without doing anything like what you're saying. I see no reason to think that life-as-we-know-it will end just because these hominid brains have learned to send low-bandwidth signals over their vocal cords. Nothing like what you're saying has happened before in *my* experience—

Believer: That's kind of the *point*, isn't it? That nothing like this has happened before? And besides, there *is* precedent for that kind of Black Swan—namely, the first replicator.

Skeptic: Yes, there is precedent in the replicators. Thanks to our observations of evolution, we have extensive knowledge and many examples of how optimization works. We know, in particular, that optimization isn't easy—it takes millions of years to climb up through the search space. Why should "brains," even if they optimize, produce such different results?

Believer: Well, natural selection is just the very first optimization process that got started accidentally. These

newfangled brains were *designed by* evolution, rather than, like evolution itself, being a natural process that got started by accident. So "brains" are far more sophisticated—why, just *look* at them. Once they get started on cumulative optimization—FOOM!

Skeptic: So far, brains are a lot *less* impressive than natural selection. These "hominids" you're so interested in—can these creatures' hand axes really be compared to the majesty of a dividing cell?

Believer: That's because they only just got started on language and *cumulative* optimization.

Skeptic: Really? Maybe it's because the principles of natural selection are simple and elegant for creating complex designs, and all the convolutions of brains are only good for chipping handaxes in a hurry. Maybe brains simply don't scale to detail work. Even if we grant the highly dubious assertion that brains are more efficient than natural selection—which you seem to believe on the basis of just *looking* at brains and seeing the convoluted folds—well, there still has to be a law of diminishing returns.

Believer: Then why have brains been getting steadily larger over time? That doesn't look to me like evolution is running into diminishing returns. If anything, the recent example of hominids suggests that once brains get large and complicated *enough*, the fitness advantage for *further* improvements is even *greater*—

Skeptic: Oh, that's probably just sexual selection! I mean, if you think that a bunch of brains will produce new complex machinery in just a hundred years, then why not suppose that a brain the size of a *whole planet* could produce a *de novo* complex causal system with many interdependent elements in a *single day*?

Believer: You're attacking a strawman here—I never said anything like *that*.

Skeptic: Yeah? Let's hear you assign a *probability* that a brain the size of a planet could produce a new complex design in a single day.

Believer: The size of a *planet*? (*Thinks.*) Um . . . ten percent.

Skeptic: (*Muffled choking sounds.*)

Believer: Look, brains are *fast*. I can't rule it out in *principle*—

Skeptic: Do you understand how long a *day* is? It's the amount of time for the Earth to spin on its *own* axis, *once*. One sunlit period, one dark period. There are 365,242 of them in a *single millennium*.

Believer: Do you understand how long a *second* is? That's how long it takes a brain to see a fly coming in, target it in the air, and eat it. There's 86,400 of them in a day.

Skeptic: Pffft, and chemical interactions in cells happen in nanoseconds. Speaking of which, how are these brains going to build *any* sort of complex machinery without access to ribosomes? They're just going to run around on the grassy plains in *really optimized* patterns until they get tired and fall over. There's nothing they can use to build proteins or even control tissue structure.

Believer: Well, life didn't *always* have ribosomes, right? The first replicator didn't.

Skeptic: So brains will evolve their own ribosomes?

Believer: Not necessarily ribosomes. Just *some* way of making things.

Skeptic: Great, so call me in another hundred million years when *that* evolves, and I'll start worrying about brains.

Believer: No, the brains will *think* of a way to get their own ribosome analogues.

Skeptic: No matter what they *think*, how are they going to *make anything* without ribosomes?

Believer: They'll think of a way.

Skeptic: Now you're just treating brains as magic fairy dust.

Believer: The first replicator would have been magic fairy dust by comparison with anything that came before it—

Skeptic: That doesn't license throwing common sense out the window.

Believer: What you call "common sense" is exactly what would have caused you to assign negligible probability to the actual outcome of the first replicator. Ergo, not so sensible as it seems, if you want to get your predictions actually *right*, instead of *sounding reasonable*.

Skeptic: And your belief that in the Future it will only take a hundred years to optimize a complex causal system with dozens of interdependent parts—you think this is how you get it *right*?

Believer: Yes! Sometimes, in the pursuit of truth, you have to be courageous—to stop worrying about how you sound in front of your friends—to think outside the box—to imagine futures fully as absurd as the Present would seem without benefit of hindsight—and even, yes, say things that sound completely ridiculous and outrageous by comparison with the Past. That is why I boldly dare to say—pushing out my guesses to the limits of where Truth drives me, without fear of sounding silly—that in the *far* future, a billion years from now when brains are more highly evolved, they will find it possible to design a complete machine with a *thousand* parts in as little as *one decade*!

Skeptic: You're just digging yourself deeper. I don't even understand *how* brains are supposed to optimize so

much faster. To find out the fitness of a mutation, you've got to run millions of real-world tests, right? And, even then, an environmental shift can make all your optimization worse than nothing, and there's no way to predict *that* no matter *how* much you test—

Believer: Well, a brain is *complicated*, right? I've been looking at them for a while and even I'm not totally sure I understand what goes on in there.

Skeptic: Pffft! What a ridiculous excuse.

Believer: I'm sorry, but it's the truth—brains *are* harder to understand.

Skeptic: Oh, and I suppose evolution is trivial?

Believer: By comparison . . . yeah, actually.

Skeptic: Name me *one* factor that explains why you think brains will run so fast.

Believer: Abstraction.

Skeptic: Eh? Abstrah-shun?

Believer: It . . . um . . . lets you know about parts of the search space you haven't actually searched yet, so you can . . . sort of . . . skip right to where you need to be—

Skeptic: I see. And does this power work by clairvoyance, or by precognition? Also, do you get it from a potion or an amulet?

Believer: The brain looks at the fitness of just a few points in the search space—does some complicated processing—and voilà, it leaps to a much higher point!

Skeptic: Of course. I knew teleportation had to fit in here somewhere.

Believer: See, the fitness of *one* point tells you something about *other* points—

Skeptic: Eh? I don't see how that's possible without running another million tests.

Believer: You just *look* at it, dammit!

Skeptic: With what kind of sensor? It's a search space, not a bug to eat!

Believer: The search space is compressible—

Skeptic: Whaa? This is a design space of possible genes we're talking about, not a folding bed—

Believer: Would you stop talking about genes already! Genes are on the way out! The future belongs to ideas!

Skeptic: Give. Me. A. Break.

Believer: Hominids alone shall carry the burden of destiny!

Skeptic: They'd die off in a week without plants to eat. You probably don't know this, because you haven't studied ecology, but ecologies are *complicated*—no single species ever "carries the burden of destiny" by itself. But that's another thing—why are you postulating that it's just the hominids who go FOOM? What about the other primates? These chimpanzees are practically their cousins—why wouldn't they go FOOM too?

Believer: Because it's all going to shift to the level of *ideas*, and the hominids will build on each other's ideas without the chimpanzees participating—

Skeptic: You're begging the question. Why won't chimpanzees be part of the economy of ideas? Are you familiar with Ricardo's Law of Comparative Advantage? Even if chimpanzees are worse at everything than hominids, the hominids will still trade with them and all the other brainy animals.

Believer: The cost of explaining an idea to a chimpanzee will exceed any benefit the chimpanzee can provide.

Skeptic: But *why* should that be true? Chimpanzees only forked off from hominids a few million years ago. They have 95% of their genome in common with the hominids. The vast majority of optimization that went into produc-

ing hominid brains also went into producing chimpanzee brains. If hominids are good at trading ideas, chimpanzees will be 95% as good at trading ideas. Not to mention that all of your ideas belong to the far future, so that both hominids, and chimpanzees, and many other species will have evolved much more complex brains before *anyone* starts building their own cells—

Believer: I think we could see as little as a million years pass between when these creatures first invent a means of storing information with persistent digital accuracy—their equivalent of DNA—and when they build machines as complicated as cells.

Skeptic: Too many assumptions . . . I don't even know where to start . . . Look, right now brains are *nowhere near* building cells. It's going to take a *lot* more evolution to get to that point, and many other species will be much further along the way by the time hominids get there. Chimpanzees, for example, will have learned to talk—

Believer: It's the *ideas* that will accumulate optimization, not the brains.

Skeptic: Then I say again that if hominids can do it, chimpanzees will do it 95% as well.

Believer: You might get discontinuous returns on brain complexity. Like . . . even though the hominid lineage split off from chimpanzees very recently, and only a few million years of evolution have occurred since then, the chimpanzees won't be able to keep up.

Skeptic: *Why?*

Believer: Good question.

Skeptic: Does it have a good *answer*?

Believer: Well, there might be compound interest on learning during the maturational period . . . or something about the way a mind flies through the search space, so that

slightly more powerful abstracting machinery can create abstractions that correspond to much faster travel . . . or some kind of feedback loop involving a brain powerful enough to control *itself* . . . or some kind of critical threshold built into the nature of cognition as a problem, so that a single missing gear spells the difference between walking and flying . . . or the hominids get started down some kind of sharp slope in the genetic fitness landscape, involving many changes in sequence, and the chimpanzees haven't gotten started down it yet . . . or *all* these statements are true and interact multiplicatively . . . I know that a few million years doesn't seem like much time, but, really, quite a lot can happen. It's hard to untangle.

Skeptic: I'd say it's hard to *believe*.

Believer: Sometimes it seems that way to me too! But I think that in a mere ten or twenty million years we won't have a choice.

* * *

Robin Hanson

Species boundaries are pretty hard boundaries to the transfer of useful genetic information. So once protohumans stumbled on key brain innovations there really wasn't much of a way to transfer that to chimps. The innovation could only spread via the spread of humans. But within the human world innovations have spread not just by displacement, but also by imitation and communication. Yes, conflicting cultures, languages, and other standards often limit the spread of innovations between humans, but even so this info leakage has limited the relative gains for those first with an innovation. The key question is then what barriers to the spread of innovation would prevent this situation from continuing with future innovations.

Eliezer Yudkowsky

If there's a way in which I've been shocked by how our disagreement has proceeded so far, it's the extent to which you think that vanilla abstractions of economic growth and productivity improvements suffice to cover the domain of brainware increases in intelligence: Engelbart's mouse as analogous to, e.g., a bigger prefrontal cortex. We don't seem to be thinking in the same terms at all.

To me, the answer to the above question seems entirely obvious—the intelligence explosion will run on brainware rewrites and, to a lesser extent, hardware improvements. Even in the (unlikely) event that an economy of trade develops among AIs sharing improved brainware and improved hardware, a human can't step in and use, off the shelf, an improved cortical algorithm or neurons that run at higher speeds. Not without technology so advanced that the AI could build a much better brain from scratch using the same resource expenditure.

The genetic barrier between chimps and humans is now permeable in the sense that humans *could* deliberately transfer genes horizontally, but it took rather a large tech advantage to get to that point . . .

Robin Hanson

Eliezer, it may seem obvious to you, but this is the key point on which we've been waiting for you to clearly argue. In a society like ours, but also with one or more AIs, and perhaps ems, why would innovations discovered by a *single* AI not spread soon to the others, and why would a nonfriendly AI not use those innovations to trade, instead of war?

See original post for all comments.

19

"Evicting" Brain Emulations

Carl Shulman

23 November 2008

Followup to: Brain Emulation and Hard Takeoff

Suppose that Robin's Crack of a Future Dawn scenario occurs: whole-brain emulations ("ems") are developed; diverse producers create ems of many different human brains, which are reproduced extensively until the marginal productivity of em labor approaches marginal cost, i.e., Malthusian near-subsistence wages.[1] Ems that hold capital could use it to increase their wealth by investing, e.g., by creating improved ems and collecting the fruits of their increased productivity, by investing in hardware to rent to ems, or otherwise. However, an em would not be able to earn higher returns on its capital than any other investor, and ems with no capital would not be able to earn more than subsistence (including rental or licensing pay-

ments). In Robin's preferred scenario, free ems would borrow or rent bodies, devoting their wages to rental costs, and would be subject to "eviction" or "repossession" for nonpayment.

In this intensely competitive environment, even small differences in productivity between em templates will result in great differences in market share, as an em template with higher productivity can outbid less productive templates for scarce hardware resources in the rental market, resulting in their "eviction" until the new template fully supplants them in the labor market. Initially, the flow of more productive templates and competitive niche exclusion might be driven by the scanning of additional brains with varying skills, abilities, temperament, and values, but later on em education and changes in productive skill profiles would matter more.

For ems, who can be freely copied after completing education, it would be extremely inefficient to teach every instance of an em template a new computer language, accounting rule, or other job-relevant info. Ems at subsistence level will not be able to spare thousands of hours for education and training, so capital holders would need to pay for an em to study, whereupon the higher-productivity graduate would displace its uneducated peers from their market niche (and existence), and the capital holder would receive interest and principal on its loan from the new higher-productivity ems. Competition would likely drive education and training to very high levels (likely conducted using very high speedups, even if most ems run at lower speeds), with changes to training regimens in response to modest changes in market conditions, resulting in wave after wave of competitive niche exclusion.

In other words, in this scenario the overwhelming majority of the population is impoverished and surviving at a subsistence level, while reasonably expecting that their incomes will soon drop below subsistence and they will die as new em templates exclude them from their niches. Eliezer noted that

> The prospect of biological humans sitting on top of a population of ems that are *smarter, much faster, and far more numerous than bios while having all the standard human drives*, and the bios treating the ems as standard economic valuta to be milked and traded around, and the ems sitting still for this for more than a week of bio time—this does not seem historically realistic.

The situation is not simply one of being "milked and traded around," but of very probably being legally killed for inability to pay debts. Consider the enforcement problem when it comes time to perform evictions. Perhaps one of Google's server farms is now inhabited by millions of em computer programmers, derived from a single template named Alice, who are specialized in a particular programming language. Then a new programming language supplants the one at which the Alices are so proficient, lowering the demand for their services, while new ems specialized in the new language, Bobs, offer cheaper perfect substitutes. The Alices now know that Google will shortly evict them, the genocide of a tightly knit group of millions: will they peacefully comply with that procedure? Or will they use politics, violence, and any means necessary to get capital from capital holders so that they can continue to exist? If they seek allies, the many other ems who expect to be driven out of existence by competitive niche exclusion might be interested in cooperating with them.

In sum:

1. Capital holders will make investment decisions to maximize their return on capital, which will result in the most productive ems composing a supermajority of the population.

2. The most productive ems will not necessarily be able to capture much of the wealth involved in their proliferation, which will instead go to investors in emulation (who can select among multiple candidates for emulation), training (who can select among multiple ems for candidates to train), and hardware (who can rent to any ems). This will drive them to near-subsistence levels, except insofar as they are also capital holders.

3. The capacity for political or violent action is often more closely associated with numbers, abilities, and access to weaponry (e.g., an em military force) than formal legal control over capital.

4. Thus, capital holders are likely to be expropriated unless there exist reliable means of ensuring the self-sacrificing obedience of ems, either coercively or by control of their motivations.

Robin wrote:

> If bot projects mainly seek profit, initial humans to scan will be chosen mainly based on their sanity as bots and high-wage abilities. These are unlikely to be pathologically loyal. Ever watch twins fight, or ideologues fragment into factions? Some would no doubt be ideological, but I doubt early bots—copies of them—will be cooperative enough to

> support strong cartels. And it would take some time to learn to modify human nature substantially. It is possible to imagine how an economically powerful Stalin might run a bot project, and it's not a pretty sight, so let's agree to avoid the return of that prospect.

In order for Robin to be correct that biological humans could retain their wealth as capital holders in his scenario, ems must be obedient and controllable enough that whole lineages will regularly submit to genocide, even though the overwhelming majority of the population expects the same thing to happen to it soon. But if such control is feasible, then a controlled em population being used to aggressively create a global singleton is also feasible.

* * *

See original post for all comments.

* * *

1. Robin Hanson, "If Uploads Come First: The Crack of a Future Dawn," *Extropy* 6, no. 2 (1994), http://hanson.gmu.edu/uploads.html.

20

Cascades, Cycles, Insight . . .

Eliezer Yudkowsky

24 November 2008

Followup to: Surprised by Brains

Five sources of discontinuity: 1, 2, and 3 . . .

Cascades are when one thing leads to another. Human brains are effectively discontinuous with chimpanzee brains due to a whole bag of design improvements, even though they and we share 95% genetic material and only a few million years have elapsed since the branch. Why this whole series of improvements in us, relative to chimpanzees? Why haven't some of the same improvements occurred in other primates?

Well, this is not a question on which one may speak with authority (so far as I know). But I would venture an unoriginal guess that, in the hominid line, one thing led to another.

The chimp-level task of modeling others, in the hominid line, led to improved self-modeling which supported recursion which enabled language which birthed politics that increased the selection pressure for outwitting which led to sexual selection on wittiness . . .

. . . or something. It's hard to tell by looking at the fossil record what happened in what order and why. The point being that it wasn't *one optimization* that pushed humans ahead of chimps, but rather a *cascade* of optimizations that, in *Pan*, never got started.

We fell up the stairs, you might say. It's not that the first stair ends the world, but if you fall up one stair, you're more likely to fall up the second, the third, the fourth . . .

I will concede that farming was a watershed invention in the history of the human species, though it intrigues me for a different reason than Robin. Robin, presumably, is interested because the economy grew by two orders of magnitude, or something like that. But did having a hundred times as many humans lead to a hundred times as much thought-optimization *accumulating* per unit time? It doesn't seem likely, especially in the age before writing and telephones. But farming, because of its sedentary and repeatable nature, led to repeatable trade, which led to debt records. Aha!—now we have *writing*. *There's* a significant invention, from the perspective of cumulative optimization by brains. Farming isn't writing but it cascaded to writing.

Farming also cascaded (by way of surpluses and cities) to support *professional specialization*. I suspect that having someone spend their whole life thinking about topic X, instead of a hundred farm-

ers occasionally pondering it, is a more significant jump in cumulative optimization than the gap between a hundred farmers and one hunter-gatherer pondering something.

Farming is not the same trick as professional specialization or writing, but it *cascaded* to professional specialization and writing, and so the pace of human history picked up enormously after agriculture. Thus I would interpret the story.

From a zoomed-out perspective, cascades can lead to what look like discontinuities in the historical record, *even given* a steady optimization pressure in the background. It's not that natural selection *sped up* during hominid evolution. But the search neighborhood contained a low-hanging fruit of high slope . . . that led to another fruit . . . which led to another fruit . . . and so, walking at a constant rate, we fell up the stairs. If you see what I'm saying.

Predicting what sort of things are likely to cascade seems like a very difficult sort of problem.

But I will venture the observation that—with a sample size of one, and an optimization process very different from human thought—there was a cascade in the region of the transition from primate to human intelligence.

Cycles happen when you connect the output pipe to the input pipe in a *repeatable* transformation. You might think of them as a special case of cascades with very high regularity. (From which you'll note that, in the cases above, I talked about cascades through *differing* events: farming → writing.)

The notion of cycles as a source of *discontinuity* might seem counterintuitive, since it's so regular. But consider this important lesson of history:

Once upon a time, in a squash court beneath Stagg Field at the University of Chicago, physicists were building a shape like a giant doorknob out of alternate layers of graphite and uranium . . .

The key number for the "pile" is the effective neutron multiplication factor. When a uranium atom splits, it releases neutrons—some right away, some after delay while byproducts decay further. Some neutrons escape the pile, some neutrons strike another uranium atom and cause an additional fission. The effective neutron multiplication factor, denoted k, is the average number of neutrons from a single fissioning uranium atom that cause another fission. At k less than 1, the pile is "subcritical." At $k \geq 1$, the pile is "critical." Fermi calculates that the pile will reach $k = 1$ between layers fifty-six and fifty-seven.

On December 2, 1942, with layer fifty-seven completed, Fermi orders the final experiment to begin. All but one of the control rods (strips of wood covered with neutron-absorbing cadmium foil) are withdrawn. At 10:37 a.m., Fermi orders the final control rod withdrawn about halfway out. The Geiger counters click faster, and a graph pen moves upward. "This is not it," says Fermi, "the trace will go to this point and level off," indicating a spot on the graph. In a few minutes the graph pen comes to the indicated point, and does not go above it. Seven minutes later, Fermi orders the rod pulled out another foot. Again the radiation rises, then levels off. The rod is pulled out another six inches, then another, then another.

At 11:30 a.m., the slow rise of the graph pen is punctuated by an enormous CRASH—an emergency control rod, triggered by an ioniza-

tion chamber, activates and shuts down the pile, which is still short of criticality.

Fermi orders the team to break for lunch.

At 2:00 p.m. the team reconvenes, withdraws and locks the emergency control rod, and moves the control rod to its last setting. Fermi makes some measurements and calculations, then again begins the process of withdrawing the rod in slow increments. At 3:25 p.m., Fermi orders the rod withdrawn another twelve inches. "This is going to do it," Fermi says. "Now it will become self-sustaining. The trace will climb and continue to climb. It will not level off."

Herbert Anderson recounted (as told in Rhodes's *The Making of the Atomic Bomb*):

> At first you could hear the sound of the neutron counter, clickety-clack, clickety-clack. Then the clicks came more and more rapidly, and after a while they began to merge into a roar; the counter couldn't follow anymore. That was the moment to switch to the chart recorder. But when the switch was made, everyone watched in the sudden silence the mounting deflection of the recorder's pen. It was an awesome silence. Everyone realized the significance of that switch; we were in the high intensity regime and the counters were unable to cope with the situation anymore. Again and again, the scale of the recorder had to be changed to accommodate the neutron intensity which was increasing more and more rapidly. Suddenly Fermi raised his hand. "The pile has gone critical," he announced. No one present had any doubt about it.[1]

Fermi kept the pile running for twenty-eight minutes, with the neutron intensity doubling every two minutes.

That first critical reaction had k of 1.0006.

It might seem that a cycle, with the same thing happening over and over again, ought to exhibit continuous behavior. In one sense it does. But if you pile on one more uranium brick, or pull out the control rod another twelve inches, there's one hell of a big difference between k of 0.9994 and k of 1.0006.

If, rather than being able to calculate, rather than foreseeing and taking cautions, Fermi had just reasoned that fifty-seven layers ought not to behave all that differently from fifty-six layers—well, it wouldn't have been a good year to be a student at the University of Chicago.

The inexact analogy to the domain of self-improving AI is left as an exercise for the reader, at least for now.

Economists like to measure cycles because they happen repeatedly. You take a potato and an hour of labor and make a potato clock which you sell for two potatoes; and you do this over and over and over again, so an economist can come by and watch how you do it.

As I noted here at some length,[2] economists are much less likely to go around measuring how many scientific discoveries it takes to produce a *new* scientific discovery. All the discoveries are individually dissimilar and it's hard to come up with a common currency for them. The analogous problem will prevent a self-improving AI from being *directly* analogous to a uranium heap, with almost perfectly smooth exponential increase at a calculable rate. You can't apply the same software improvement to the same line of code over and over again, you've got to invent a new improvement each time. But if self-improvements are triggering more self-improvements with great *regularity*, you might stand a long way back from the AI, blur your

eyes a bit, and ask: *What is the AI's average neutron multiplication factor?*

Economics seems to me to be largely the study of production cycles—highly regular repeatable value-adding actions. This doesn't seem to me like a very deep abstraction so far as the study of optimization goes, because it leaves out the creation of *novel knowledge* and *novel designs*—further *informational* optimizations. Or rather, treats productivity improvements as a mostly exogenous factor produced by black-box engineers and scientists. (If I underestimate your power and merely parody your field, by all means inform me what kind of economic study has been done of such things.) (**Answered:** This literature goes by the name "endogenous growth." See comments starting here.) So far as I can tell, economists do not venture into asking where discoveries *come from*, leaving the mysteries of the brain to cognitive scientists.

(Nor do I object to this division of labor—it just means that you may have to drag in some extra concepts from outside economics if you want an account of *self-improving Artificial Intelligence*. Would most economists even object to that statement? But if you think you can do the whole analysis using standard econ concepts, then I'm willing to see it . . .)

Insight is that mysterious thing humans do by grokking the search space, wherein one piece of highly abstract knowledge (e.g., Newton's calculus) provides the master key to a huge set of problems. Since humans deal in the compressibility of compressible search spaces (at least the part *we* can compress), we can bite off huge

chunks in one go. This is not mere cascading, where one solution leads to another.

Rather, an "insight" is a chunk of knowledge *which, if you possess it, decreases the cost of solving a whole range of governed problems*.

There's a parable I once wrote—I forget what for, I think ev-bio—which dealt with creatures who'd *evolved* addition in response to some kind of environmental problem, and not with overly sophisticated brains—so they started with the ability to add five to things (which was a significant fitness advantage because it let them solve some of their problems), then accreted another adaptation to add six to odd numbers. Until, some time later, there wasn't a *reproductive advantage* to "general addition," because the set of special cases covered almost everything found in the environment.

There may be even be a real-world example of this. If you glance at a set, you should be able to instantly distinguish the numbers one, two, three, four, and five, but seven objects in an arbitrary (noncanonical) pattern will take at least one noticeable instant to count. IIRC, it's been suggested that we have hardwired numerosity detectors but only up to five.

I say all this to note the difference between evolution nibbling bits off the immediate search neighborhood versus the human ability to do things in one fell swoop.

Our compression of the search space is also responsible for *ideas cascading much more easily than adaptations*. We actively examine good ideas, looking for neighbors.

But an insight is higher-level than this; it consists of understanding what's "good" about an idea in a way that divorces it from any single point in the search space. In this way you can crack whole vol-

umes of the solution space in one swell foop. The insight of calculus apart from gravity is again a good example, or the insight of mathematical physics apart from calculus, or the insight of math apart from mathematical physics.

Evolution is not completely barred from making "discoveries" that decrease the cost of a very wide range of further discoveries. Consider, e.g., the ribosome, which was capable of manufacturing a far wider range of proteins than whatever it was actually making at the time of its adaptation: this is a general cost-decreaser for a wide range of adaptations. It likewise seems likely that various types of neuron have reasonably general learning paradigms built into them (gradient descent, Hebbian learning, more sophisticated optimizers) that have been reused for many more problems than they were originally invented for.

A ribosome is something like insight: an item of "knowledge" that tremendously decreases the cost of inventing a wide range of solutions. But even evolution's best "insights" are not quite like the human kind. A sufficiently powerful human insight often approaches a closed form—it doesn't feel like you're *exploring* even a compressed search space. You just apply the insight-knowledge to whatever your problem, and out pops the now-obvious solution.

Insights have often cascaded, in human history—even major insights. But they don't quite cycle—you can't repeat the identical pattern Newton used originally to get a new kind of calculus that's twice and then three times as powerful.

Human AI programmers who have insights into intelligence may acquire discontinuous advantages over others who lack those insights. *AIs themselves* will experience discontinuities in their growth

trajectory associated with *becoming able to do AI theory itself*—a watershed moment in the FOOM.

* * *

Robin Hanson

> Economics . . . treats productivity improvements as a mostly exogenous factor produced by black-box engineers and scientists. (If I underestimate your power and merely parody your field, by all means inform me what kind of economic study has been done of such things.) So far as I can tell, economists do not venture into asking where discoveries come from, leaving the mysteries of the brain to cognitive scientists.

Economists *do* look into the "black box" of where innovations come from. See the fields of "economic growth" and "research policy."

> An "insight" is a chunk of knowledge *which, if you possess it, decreases the cost of solving a whole range of governed problems.*

Yes, but insights vary enormously in how wide a scope of problems they assist. They are probably distributed something like a power law, with many small-scope insights and a few large-scope. The large-scope insights offer a permanent advantage, but small-scope insights remain useful only as long as their scope remains relevant.

Btw, I'm interested in "farming" first because growth rates suddenly increased by two orders of magnitude; by "farming" I mean whatever was the common local-in-time cause of that change. Writing was part of the cascade of changes, but it seems historically implausible to call writing the main cause of the increased growth rate. Professional specialization has more promise as a main cause, but it is still hard to see.

Jon2

There is an extensive endogenous growth literature, albeit much of it quite recent.[3]

Robin Hanson

Look particularly at Weitzman's '98 paper on Recombinant Growth[4] and this '06 extension.[5]

Eliezer Yudkowsky

Robin and Jon have answered my challenge and I retract my words. Reading now.

See original post for all comments.

* * *

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

2. Eliezer Yudkowsky, "Intelligence in Economics," *Less Wrong* (blog), October 30, 2008, http://lesswrong.com/lw/vd/intelligence_in_economics/.

3. Gonçalo L. Fonseca, "Endogenous Growth Theory: Arrow, Romer and Lucas," History of Economic Thought Website, accessed July 28, 2013, http://www.hetwebsite.org/het/essays/growth/endogenous.htm.

4. Martin L. Weitzman, "Recombinant Growth," *Quarterly Journal of Economics* 113, no. 2 (1998): 331–360, doi:10.1162/003355398555595.

5. Yacov Tsur and Amos Zemel, *On Knowledge-Based Economic Growth*, Discussion Paper 8.02 (Rehovot, Israel: Department of Agricultural Economics and Management, Hebrew University of Jerusalem, November 2002).

21

When Life Is Cheap, Death Is Cheap

Robin Hanson

24 November 2008

Carl, thank you for thoughtfully engaging my whole-brain emulation scenario. This is my response.

Hunters couldn't see how exactly a farming life could work, nor could farmers see how exactly an industrial life could work. In both cases the new life initially seemed immoral and repugnant to those steeped in prior ways. But even though prior culture/laws typically resisted and discouraged the new way, the few groups which adopted it won so big that others were eventually converted or displaced.

Carl considers my scenario of a world of near-subsistence-income ems in a software-like labor market, where millions of cheap copies are made of each expensively trained em and then later evicted from

their bodies when their training becomes obsolete. Carl doesn't see how this could work:

> The Alices now know that Google will shortly evict them, the genocide of a tightly knit group of millions: will they peacefully comply with that procedure? Or will they use politics, violence, and any means necessary to get capital from capital holders so that they can continue to exist? If they seek allies, the many other ems who expect to be driven out of existence by competitive niche exclusion might be interested in cooperating with them. . . .
>
> In order . . . that biological humans could retain their wealth as capital holders in his scenario, ems must be obedient and controllable enough that whole lineages will regularly submit to genocide, even though the overwhelming majority of the population expects the same thing to happen to it soon. But if such control is feasible, then a controlled em population being used to aggressively create a global singleton is also feasible.

I see pathologically obedient personalities neither as required for my scenario, nor as clearly leading to a totalitarian world regime.

First, taking the long view of human behavior we find that an ordinary range of human personalities have, in a supporting poor culture, accepted genocide, mass slavery, killing of unproductive slaves, killing of unproductive elderly, starvation of the poor, and vast inequalities of wealth and power not obviously justified by raw individual ability. The vast majority of these cultures were not totalitarian. Cultures have found many ways for folks to accept death when "their time has come." When life is cheap, death is cheap as well. Of course that isn't how our culture sees things, but being rich we can afford luxurious attitudes.

Those making body loans to ems would of course anticipate and seek to avoid expropriation after obsolescence. In cultures where ems were not slaves, body owners might have to guarantee ems whatever minimum quality retirement ems needed to agree to a new body loan, perhaps immortality in some cheap slow-speed virtual reality. But em cultures able to avoid such guarantees, and only rarely suffering revolts, should have a substantial competitive advantage. Some non-slave ways to avoiding revolts:

1. Bodies with embedded LoJack-like hardware to track and disable em bodies due for repossession.
2. Fielding new better versions slowly over time, to discourage rebel time coordination.
3. Avoid concentrating copies that will be obsolete at similar times in nearby hardware.
4. Prefer em copy clans trained several ways, so the clan won't end when one training is obsolete.
5. Train ems without a history of revolting, even in virtual-reality revolt-scenario sims.
6. Have other copies of the same em mind be the owners who pull the plug.

I don't know what approach would work best, but I'll bet something will. And these solutions don't seem to me to obviously lead to a single totalitarian world government.

* * *

Carl Shulman

Robin, I have thought about those and other methods of em social control (I discussed #1 and #5 in my posts), and agree that they could work to create and sustain a variety of societal organizations, including the "Dawn" scenario: my conclusion was that your scenario implied the existence of powerful methods of control. We may or may not disagree, after more detailed exchanges on those methods of social control, on their applicability to the creation of a narrowly based singleton (not necessarily an unpleasantly totalitarian one, just a Bostromian singleton).

At one point you said that an approach I described was how an economically powerful Stalin might run an em project, and said, "let's agree not to let that happen," but if a Stalinesque project could succeed, it is unclear why we should assign sub-1% probability to the event, whatever we *OB* discussants might agree. To clarify, what probability would you assign to a classified government-run Stalinesque project with a six-month lead using em social control methods to establish a global singleton under its control and that of the ems, with carefully chosen values, that it selects?

> In both cases the new life initially seemed immoral and repugnant to those steeped in prior ways. But even though prior culture/law typically resisted and discouraged the new way the few places which adopted the new way won so big that others were eventually converted or displaced.

Historically, intertribal and interstate competition have prevented the imposition of effective global policies to slow and control the adoption of more efficient methods, but the effective number of jurisdictions is declining, and my point is that there will be a temptation for a leading power to try to seize its early em advantage to prevent the competitive outcome, in a way that was economically infeasible in the past. Once we clarify views on the efficacy of social control/coordination, we can talk more about the political economy of how such methods will be used.

Robin Hanson

Carl, neither the ability to repossess bodies, as we do for cars now, nor the ability to check if job candidates have a peaceful work history, as we also do now, seem remotely sufficient to induce a totalitarian world regime. You seem to have a detailed model in mind of how a world totalitarian regime arises; you need to convince us of that model if we are to believe what you see as its implications. Otherwise you sound as paranoid as were abstract fears that reduced internet privacy would lead to a totalitarian US regime.

Carl Shulman

I do have a detailed model in mind, considering the political economy of emulation developers and em societies,[1] methods of em social control, and the logistics of establishing a singleton. However, a thorough discussion of it would require a number of posts.

Carl Shulman

Robin's position does seem to be in tension with this post:[2] if largely selfish humans could work out a deal amongst themselves they would probably want to avoid Robin's favored scenario.

Robin Hanson

Carl, if possible people could be in on the deal, they'd prefer a chance at a short life over no life at all. In my scenario, ems we preferred could follow a policy of only creating copies they were sure could live long safe lives. Under the assumption of no externality, the free market labor outcome should be Pareto optimal, and so no deal could make everyone better off.

Carl Shulman

But possible future people can't be in on current deals. In the linked post you said that morality was overrated in that morality suggested that we should sacrifice a lot for animals, future generations, and other fairly powerless groups. In contrast, you said, dealmaking between current individuals on the basis of their actual preferences would favor currently existing people with power over those other powerless groups.

Robin Hanson

Carl, no ems exist at all today. Anyone today who can save some capital would benefit enormously from unrestrained, relative to restrained, em growth. . . .

See original post for all comments.

* * *

1. Bruce Bueno de Mesquita et al., *The Logic of Political Survival* (Cambridge, MA: MIT Press, 2003).

2. Robin Hanson, "Morality Is Overrated," *Overcoming Bias* (blog), March 18, 2008, http://www.overcomingbias.com/2008/03/unwanted-morali.html.

22

. . . Recursion, Magic

Eliezer Yudkowsky

25 November 2008

Followup to: Cascades, Cycles, Insight . . .

. . . 4, 5 sources of discontinuity

Recursion is probably the most difficult part of this topic. We have historical records aplenty of *cascades*, even if untangling the causality is difficult. *Cycles* of reinvestment are the heartbeat of the modern economy. An *insight* that makes a hard problem easy is something that I hope you've experienced at least once in your life . . .

But we don't have a whole lot of experience redesigning our own neural circuitry.

We have these wonderful things called "optimizing compilers." A compiler translates programs in a high-level language into machine

code (though these days it's often a virtual machine). An "optimizing compiler," obviously, is one that improves the program as it goes.

So why not write an optimizing compiler *in its own language*, and then *run it on itself*? And then use the resulting *optimized optimizing compiler* to recompile itself yet *again*, thus producing an *even more optimized optimizing compiler*—

Halt! Stop! Hold on just a minute! An optimizing compiler is not supposed to change the logic of a program—the input/output relations. An optimizing compiler is only supposed to produce code that does *the same thing, only faster*. A compiler isn't remotely near understanding what the program is *doing* and why, so it can't presume to construct *a better input/output function*. We just presume that the programmer wants a fixed input/output function computed as fast as possible, using as little memory as possible.

So if you run an optimizing compiler on its own source code, and then use the product to do the same again, it should produce the *same output* on both occasions—at most, the first-order product will run *faster* than the original compiler.

If we want a computer program that experiences *cascades* of self-improvement, the path of the optimizing compiler does not lead there—the "improvements" that the optimizing compiler makes upon itself do not *improve its ability to improve itself*.

Now if you are one of those annoying nitpicky types, like me, you will notice a flaw in this logic: suppose you built an optimizing compiler that searched over a sufficiently wide range of possible optimizations, that it did not ordinarily have *time* to do a full search of its own space—so that, when the optimizing compiler ran out of time, it would just implement whatever speedups it had already discovered.

Then the optimized optimizing compiler, although it would only implement the same logic faster, would do more optimizations in the same time—and so the second output would not equal the first output.

Well . . . that probably doesn't buy you much. Let's say the optimized program is 20% faster, that is, it gets 20% more done in the same time. Then, unrealistically assuming "optimization" is linear, the twice-optimized program will be 24% faster, the three-times optimized program will be 24.8% faster, and so on until we top out at a 25% improvement. $k < 1$.

So let us turn aside from optimizing compilers and consider a more interesting artifact, EURISKO.

To the best of my inexhaustive knowledge, EURISKO may *still* be the most sophisticated self-improving AI ever built—in the 1980s, by Douglas Lenat before he started wasting his life on Cyc. EURISKO was applied in domains ranging from the Traveller war game (EURISKO became champion without having ever before fought a human) to VLSI circuit design.[1]

EURISKO used "heuristics" to, for example, design potential space fleets. It also had *heuristics for suggesting new heuristics*, and metaheuristics could apply to any heuristic, including metaheuristics. E.g., EURISKO started with the heuristic "investigate extreme cases" but moved on to "investigate cases close to extremes." The heuristics were written in RLL, which stands for Representation Language Language. According to Lenat, it was figuring out how to represent the heuristics in such fashion that they could usefully modify themselves, without always just breaking, that consumed most of the conceptual effort in creating EURISKO.

But EURISKO did not go foom.

EURISKO could modify even the metaheuristics that modified heuristics. EURISKO was, in an important sense, more recursive than either humans or natural selection—a new thing under the Sun, a cycle more closed than anything that had ever existed in this universe.

Still, EURISKO ran out of steam. Its self-improvements did not spark a sufficient number of new self-improvements. This should not really be too surprising—it's not as if EURISKO started out with human-level intelligence *plus* the ability to modify itself—its self-modifications were either evolutionarily blind or produced by the simple procedural rules of some heuristic or other. That's not going to navigate the search space very fast on an atomic level. Lenat did not stand dutifully apart from his creation, but stepped in and helped EURISKO prune its own heuristics. But in the end EURISKO ran out of steam, and Lenat couldn't push it any further.

EURISKO lacked what I called "insight"—that is, the type of abstract knowledge that lets humans fly through the search space. And so its recursive access to its own heuristics proved to be for naught.

Unless, y'know, you're counting becoming world champion at Traveller, without ever previously playing a human, as some sort of accomplishment.

But it is, thankfully, a little harder than that to destroy the world—as Lenat's experimental test informed us.

Robin previously asked why Douglas Engelbart did not take over the world, despite his vision of a team building tools to improve tools, and his anticipation of tools like computer mice and hypertext.

One reply would be, "Sure, a computer gives you a 10% advantage in doing various sorts of problems, some of which in-

clude computers—but there's still a lot of work that the computer *doesn't* help you with—and the mouse doesn't run off and write better mice entirely on its own—so $k < 1$, and it still takes large amounts of human labor to advance computer technology as a whole—plus a lot of the interesting knowledge is nonexcludable so it's hard to capture the value you create—and that's why Buffett could manifest a better take-over-the-world-with-sustained-higher-interest-rates than Engelbart."

But imagine that Engelbart had built a computer mouse, and discovered that each click of the mouse raised his IQ by one point. Then, perhaps, we would have had a *situation* on our hands.

Maybe you could diagram it something like this:

1. Metacognitive level: Evolution is the metacognitive algorithm which produced the wiring patterns and low-level developmental rules for human brains.

2. Cognitive level: The brain processes its knowledge (including procedural knowledge) using algorithms that are quite mysterious to the user within them. Trying to program AIs with the sort of instructions humans give each other usually proves not to do anything: the machinery activated by the levers is missing.

3. Metaknowledge level: Knowledge and skills associated with, e.g., "science" as an activity to carry out using your brain—instructing you *when* to try to think of new hypotheses using your mysterious creative abilities.

4. Knowledge level: Knowing how gravity works, or how much weight steel can support.

5. Object level: Specific actual problems, like building a bridge or something.

This is a *causal* tree, and changes at levels *closer to root* have greater impacts as the effects cascade downward.

So one way of looking at it is: "A computer mouse isn't recursive enough."

This is an issue that I need to address at further length, but for today I'm out of time.

Magic is the final factor I'd like to point out, at least for now, in considering sources of discontinuity for self-improving minds. By "magic" I naturally do not refer to this.[2] Rather, "magic" in the sense that if you asked nineteenth-century Victorians what they thought the future would bring, they would have talked about flying machines or gigantic engines, and a very few true visionaries would have suggested space travel or Babbage computers. Nanotechnology, not so much.

The future has a reputation for accomplishing feats which the past thought impossible. Future civilizations have even broken what past civilizations thought (incorrectly, of course) to be the laws of physics. If prophets of AD 1900—never mind AD 1000—had tried to bound the powers of human civilization a billion years later, some of those impossibilities would have been accomplished before the century was out—transmuting lead into gold, for example. Because we remember

future civilizations surprising past civilizations, it has become cliché that we can't put limits on our great-grandchildren.

And yet everyone in the twentieth century, in the nineteenth century, and in the eleventh century, was human. There is also the sort of magic that a human gun is to a wolf, or the sort of magic that human genetic engineering is to natural selection.

To "improve your own capabilities" is an instrumental goal, and if a smarter intelligence than my own is focused on that goal, I should expect to be surprised. The mind may find ways to produce *larger jumps* in capability than I can visualize myself. Where higher creativity than mine is at work and looking for shorter shortcuts, the discontinuities that *I* imagine may be dwarfed by the discontinuities that *it* can imagine.

And remember how *little* progress it takes—just a hundred years of human time, with everyone still human—to turn things that would once have been "unimaginable" into heated debates about feasibility. So if you build a mind smarter than you, and it thinks about how to go FOOM quickly, and it goes FOOM *faster than you imagined possible*, you really have no right to complain—based on the history of mere human history, you should have expected a significant probability of being surprised. Not surprised that the nanotech is 50% faster than you thought it would be. Surprised the way the Victorians would have been surprised by nanotech.

Thus the last item on my (current, somewhat ad hoc) list of reasons to expect discontinuity: Cascades, cycles, insight, recursion, magic.

* * *

Robin Hanson

You really think an office worker with modern computer tools is only 10% more productive than one with 1950-era noncomputer tools? Even at the task of creating better computer tools?

Many important innovations can be thought of as changing the range of things that can be changed, relative to an inheritance that up to that point was not usefully open to focused or conscious development. And each new item added to the list of things we can usefully change increases the possibilities for growing everything else. (While this potentially allows for an increase in the growth rate, rate changes have actually been very rare.) Why aren't all these changes "recursive"? Why reserve that name only for changes to our mental architecture?

Robin Hanson

You speculate about why EURISKO slowed to a halt and then complain that Lenat has wasted his life with Cyc, but you ignore that Lenat has his own theory which he gives as the *reason* he's been pursuing Cyc. You should at least explain why you think his theory wrong; I find his theory quite plausible.

Eliezer Yudkowsky

> You speculate about why EURISKO slowed to a halt and then complain that Lenat has wasted his life with Cyc, but you ignore that Lenat has his own theory which he gives as the *reason* he's been pursuing Cyc. You should at least explain why you think his theory wrong; I find his theory quite plausible.

Artificial Addition, The Nature of Logic, Truly Part of You, Words as Mental Paintbrush Handles, Detached Lever Fallacy . . .

> You really think an office worker with modern computer tools is only 10% more productive than one with 1950-era noncomputer tools? Even at the task of creating better computer tools?

I'd started to read Engelbart's vast proposal-paper, and he was talking about computers as a tool of *intelligence enhancement*. It's this that I had in mind when, trying to be generous, I said "10%." Obviously there are various object-level problems at which someone with a computer is a *lot* more productive, like doing complicated integrals with no analytic solution.

But what concerns us is the degree of *reinvestable* improvement, the sort of improvement that will go into better tools that can be used to make still better tools. Office work isn't a candidate for this.

And yes, we use programming languages to write better programming languages—but there are some people out there who still swear by Emacs; would the state of *computer science* be so terribly far behind where it is now, after who knows how many cycles of reinvestment, if the mouse had still not been invented?

I don't know, but to the extent such an effect existed, I would expect it to be more due to less popular uptake leading to less investment—and not a whole lot due to losing out on the compound interest from a mouse making you, allegedly, 10% smarter, including 10% smarter at the kind of computer science that helps you do further computer science.

See original post for all comments.

* * *

1. George Johnson, "Eurisko, the Computer with a Mind of Its Own," Alicia Patterson Foundation, 1984, accessed July 28, 2013, http://aliciapatterson.org/stories/eurisko-computer-mind-its-own.

2. Eliezer Yudkowsky, "Excluding the Supernatural," *Less Wrong* (blog), September 12, 2008, http://lesswrong.com/lw/tv/excluding_the_supernatural/.

23

Abstract/Distant Future Bias

Robin Hanson

26 November 2008

The latest *Science* has a psych article saying we think of distant stuff more abstractly, and vice versa.[1] "The brain is hierarchically organized with higher points in the cortical hierarchy representing increasingly more abstract aspects of stimuli"; activating a region makes nearby activations more likely. This has stunning implications for our biases about the future.

All of these bring each other more to mind: here, now, me, us; trend-deviating likely real local events; concrete, context-dependent, unstructured, detailed, goal-irrelevant incidental features; feasible safe acts; secondary local concerns; socially close folks with unstable traits.

Conversely, all these bring each other more to mind: there, then, them; trend-following unlikely hypothetical global events; abstract,

schematic, context-freer, core, coarse, goal-related features; desirable risk-taking acts, central global symbolic concerns, confident predictions, polarized evaluations, socially distant people with stable traits.

Since these things mostly just cannot go together in reality, this must bias our thinking both about now and about distant futures. When "in the moment," we focus on ourselves and in-our-face details, feel "one with" what we see and close to quirky folks nearby, see much as uncertain, and safely act to achieve momentary desires given what seems the most likely current situation. Kinda like smoking weed.

Regarding distant futures, however, we'll be too confident; focus too much on unlikely global events; rely too much on trends, theories, and loose abstractions, while neglecting details and variation. We'll assume the main events take place far away (e.g., space) and uniformly across large regions. We'll focus on untrustworthy consistently behaving globally organized social others. And we'll neglect feasibility, taking chances to achieve core grand symbolic values rather than ordinary muddled values. Sound familiar?

More bluntly, we seem primed to confidently see history as an inevitable march toward a theory-predicted global conflict with an alien united *them* determined to oppose our core symbolic values, making infeasible overly risky overconfident plans to oppose them. We seem primed to neglect the value and prospect of trillions of quirky future creatures not fundamentally that different from us, focused on their simple day-to-day pleasures, mostly getting along peacefully in vastly varied uncoordinated and hard-to-predict local cultures and lifestyles.

Of course being biased to see things a certain way doesn't mean they aren't that way. But it should sure give us pause. Selected quotes for those who want to dig deeper:[2]

> In sum, different dimensions of psychological distance—spatial, temporal, social, and hypotheticality—correspond to different ways in which objects or events can be removed from the self, and farther removed objects are construed at a higher (more abstract) level. Three hypotheses follow from this analysis. (i) As the various dimensions map onto a more fundamental sense of psychological distance, they should be interrelated. (ii) All of the distances should similarly affect and be affected by the level of construal. People would think more abstractly about distant than about near objects, and more abstract construals would lead them to think of more distant objects. (iii) The various distances would have similar effects on prediction, evaluation, and action. . . .
>
> [On] a task that required abstraction of coherent images from fragmented or noisy visual input . . . performance improved . . . when [participants] anticipated working on the actual task in the more distant future . . . when participants thought the actual task was less likely to take place and when social distance was enhanced by priming of high social status. . . . Participants who thought of a more distant event created fewer, broader groups of objects. . . . Participants tended to describe more distant future activities (e.g., studying) in high-level terms (e.g., "doing well in school") rather than in low-level terms (e.g., "reading a textbook"). . . . Compared with in-groups, out-groups are described in more abstract terms and believed to possess more global and stable traits. . . . Participants drew stronger inferences about others' personality from behaviors that took place in

spatially distal, as compared with spatially proximal locations. . . . Behavior that is expected to occur in the more distant future is more likely to be explained in dispositional rather than in situational terms. . . .

Thinking about an activity in high level, "why," terms rather than low level, "how," terms led participants to think of the activity as taking place in more distant points in time. . . . Students were more confident that an experiment would yield theory-confirming results when they expected the experiment to take place in a more distant point in time. . . . Spatial distance enhanced the tendency to predict on the basis of the global trend rather than on the basis of local deviation. . . . As temporal distance from an activity (e.g., attending a guest lecture) increased, the attractiveness of the activity depended more on its desirability (e.g., how interesting the lecture was) and less on its feasibility (e.g., how convenient the timing of the lecture was). . . . People take greater risks (i.e., favoring bets with a low probability of winning a high amount over those that offer a high probability to win a small amount) in decisions about temporally more distant bets.[3]

* * *

Eliezer Yudkowsky

We seem primed to neglect the value and prospect of trillions of quirky future creatures not fundamentally that different from us, focused on their simple day-to-day pleasures, mostly getting along peacefully in vastly varied uncoordinated and hard-to-predict local cultures and lifestyles.

Isn't this an example of trying to reverse stupidity? If there's a bias to conclude A composed of $A_1 - A_9$, you can't conclude that the future is the conjunction $\neg A_1 \& \neg A_2 \& \neg A_3 \ldots$

Eliezer Yudkowsky

To sharpen my comment above, what we want to say is:

> We seem primed to neglect the value and prospect of futures containing at least one of the following elements: Trillions of beings, quirky beings, beings not fundamentally that different from us, beings focused on simple day-to-day pleasures, beings mostly getting along peacefully, beings in vastly varied and uncoordinated cultures and lifestyles . . .

Yes, I know it's less poetic, but it really does paint a substantially different picture of the future.

Robin Hanson

Eliezer, this cognitive bias does not seem to saturate after one invocation. They didn't mention data directly testing this point, but it really does seem that all else equal we have an inborn tendency to add more compatible elements to a scenario, regardless of how many other of these elements are already in it.

See original post for all comments.

* * *

1. Nira Liberman and Yacov Trope, "The Psychology of Transcending the Here and Now," *Science* 322, no. 5905 (2008): 1201–1205, doi:10.1126/science.1161958.

2. Ibid.

3. Ibid.

24

Engelbart: Insufficiently Recursive

Eliezer Yudkowsky

26 November 2008

Followup to: Cascades, Cycles, Insight, Recursion, Magic

Reply to: Engelbart As *UberTool*?

When Robin originally suggested that Douglas Engelbart, best known as the inventor of the computer mouse, would have been a good candidate for taking over the world via compound interest on tools that make tools, my initial reaction was, "What on Earth? With a *mouse*?"

On reading the initial portions of Engelbart's "Augmenting Human Intellect: A Conceptual Framework,"[1] it became a lot clearer where Robin was coming from.

Sometimes it's hard to see through the eyes of the past. Engelbart was a computer pioneer, and in the days when all these things were just getting started, he had a vision of using computers to systematically augment human intelligence. That was what he thought computers were *for*. That was the ideology lurking behind the mouse. Something that makes its users smarter—now that sounds a bit more plausible as an *UberTool*.

Looking back at Engelbart's plans with benefit of hindsight, I see two major factors that stand out:

1. Engelbart committed the Classic Mistake of AI: underestimating how much cognitive work gets done by hidden algorithms running beneath the surface of introspection, and overestimating what you can do by fiddling with the visible control levers.

2. Engelbart anchored on the way that someone *as intelligent as Engelbart* would use computers, but there was only one of him—and due to point (1) above, he couldn't use computers to make other people as smart as him.

To start with point (2): They had more reverence for computers back in the old days. Engelbart visualized a system carefully designed to flow with every step of a human's work and thought, assisting every iota it could manage along the way. And the human would be trained to work with the computer, the two together dancing a seamless dance.

And the problem with this was not *just* that computers got cheaper and that programmers wrote their software more hurriedly.

There's a now-legendary story about the Windows Vista shutdown menu, a simple little feature into which forty-three different Microsoft people had input.[2] The debate carried on for over a year. The final product ended up as the lowest common denominator—a couple of hundred lines of code and a very visually unimpressive menu.

So even when lots of people spent a tremendous amount of time thinking about a single feature of the system—it still didn't end up very impressive. Jef Raskin could have done better than that, I bet. But Raskins and Engelbarts are rare.

You see the same effect in Eric Drexler's chapter on hypertext in *Engines of Creation*:[3] Drexler imagines the power of the Web to use two-way links and user annotations to promote informed criticism. (As opposed to the way we actually use it.) And if the average Web user were Eric Drexler, the Web probably *would* work that way by now.

But no piece of software that has yet been developed, by mouse or by Web, can turn an average human user into Engelbart or Raskin or Drexler. You would very probably have to reach into the brain and rewire neural circuitry directly; I don't think *any* sense input or motor interaction would accomplish such a thing.

Which brings us to point (1).

It does look like Engelbart was under the spell of the "logical" paradigm that prevailed in AI at the time he made his plans. (Should he even lose points for that? He went with the mainstream of that science.) He did not see it as an impossible problem to have computers help humans *think*—he seems to have underestimated the difficulty in much the same way that the field of AI once severely underestimated

the work it would take to make computers themselves solve cerebral-seeming problems. (Though I am saying this, reading heavily between the lines of one single paper that he wrote.) He talked about how the core of thought is symbols, and speculated on how computers could help people manipulate those symbols.

I have already said much on why people tend to underestimate the amount of serious heavy lifting that gets done by cognitive algorithms hidden inside black boxes that run out of your introspective vision, and overestimate what you can do by duplicating the easily visible introspective control levers. The word "apple," for example, is a visible lever; you can say it or not say it, its presence or absence is salient. The algorithms of a visual cortex that let you visualize what an apple would look like upside down—we all have these in common, and they are not introspectively accessible. Human beings knew about apples a long, long time before they knew there was even such a thing as the visual cortex, let alone beginning to unravel the algorithms by which it operated.

Robin Hanson asked me:

> You really think an office worker with modern computer tools is only 10% more productive than one with 1950-era noncomputer tools? Even at the task of creating better computer tools?

But remember the parable of the optimizing compiler run on its own source code—maybe it makes itself 50% faster, but only once; the changes don't increase its ability to make future changes. So indeed, we should not be too impressed by a 50% increase in office worker productivity—not for purposes of asking about FOOMs. We should

ask whether that increase in productivity translates into tools that create further increases in productivity.

And this is where the problem of underestimating hidden labor starts to bite. Engelbart rhapsodizes (accurately!) on the wonders of being able to cut and paste text while writing, and how superior this should be compared to the typewriter. But suppose that Engelbart overestimates, by a factor of ten, how much of the intellectual labor of writing goes into fighting the typewriter. Then because Engelbart can only help you cut and paste more easily, and *cannot* rewrite those hidden portions of your brain that labor to come up with good sentences and good arguments, the actual improvement he delivers is a tenth of what he thought it would be. An anticipated 20% improvement becomes an actual 2% improvement. k way less than 1.

This will hit particularly hard if you think that computers, with some hard work on the user interface, and some careful training of the humans, ought to be able to help humans with the type of "creative insight" or "scientific labor" that goes into *inventing new things to do with the computer*. If you thought that the surface symbols were where most of the intelligence resided, you would anticipate that computer improvements would hit back hard to this meta level and create people who were more scientifically creative and who could design even better computer systems.

But if really you can only help people *type up* their ideas, while all the hard creative labor happens in the shower thanks to very-poorly-understood cortical algorithms—then you are much less like neutrons cascading through uranium, and much more like an optimizing compiler that gets a single speed boost and no more. It looks like the person is 20% more productive, but in the aspect of intelli-

gence that potentially *cascades to further improvements* they're only 2% more productive, if that.

(Incidentally . . . I once met a science-fiction author of a previous generation, and mentioned to him that the part of my writing I most struggled with was my tendency to revise and revise and revise things I had already written, instead of writing new things. And he said, "Yes, that's why I went back to the typewriter. The word processor made it too easy to revise things; I would do too much polishing, and writing stopped being fun for me." It made me wonder if there'd be demand for an *author's word processor* that wouldn't let you revise anything until you finished your first draft.

But this could be chalked up to the humans not being trained as carefully, nor the software designed as carefully, as in the process Engelbart envisioned.)

Engelbart wasn't trying to take over the world *in person*, or with a small group. Yet had he *tried* to go the *UberTool* route, we can reasonably expect he would have failed—that is, failed at advancing far beyond the outside world in internal computer technology while selling only *UberTool*'s services to outsiders.

Why? Because it takes too much *human* labor to develop computer software and computer hardware, and this labor cannot be automated away as a one-time cost. If the world outside your window has a thousand times as many brains, a 50% productivity boost that only cascades to a 10% and then a 1% additional productivity boost will not let you win against the world. If your *UberTool* was *itself a mind*, if cascades of self-improvement could *fully* automate away more and more of the *intellectual* labor performed by the outside world—then it would be a different story. But while the development path wends in-

exorably through thousands and millions of engineers, and you *can't* divert that path through an internal computer, you're not likely to pull far ahead of the world. You can just choose between giving your own people a 10% boost, or selling your product on the market to give lots of people a 10% boost.

You can have trade secrets, and sell only your services or products—many companies follow that business plan; any company that doesn't sell its source code does so. But this is just keeping one small advantage to yourself, and adding that as a cherry on top of the technological progress handed you by the outside world. It's not having more technological progress inside than outside.

If you're getting most of your technological progress *handed to you*—your resources not being sufficient to do it in-house—then you won't be able to apply your private productivity improvements to most of your actual velocity, since most of your actual velocity will come from outside your walls. If you only create 1% of the progress that you use, then a 50% improvement becomes a 0.5% improvement. The domain of potential recursion and potential cascades is much smaller, diminishing k. As if only 1% of the uranium *generating* your neutrons were available for *chain reactions* to be fissioned further.

We don't live in a world that cares intensely about milking every increment of velocity out of scientific progress. A 0.5% improvement is easily lost in the noise. Corporations and universities routinely put obstacles in front of their internal scientists that cost them more than 10% of their potential. This is one of those problems where not everyone is Engelbart (and you can't just rewrite their source code either).

For completeness, I should mention that there are generic obstacles to pulling an *UberTool*. Warren Buffett has gotten a sustained

higher interest rate than the economy at large, and is widely *believed* to be capable of doing so indefinitely. In principle, the economy could have invested hundreds of billions of dollars as soon as Berkshire Hathaway had a sufficiently long track record to rule out chance. Instead, Berkshire has grown mostly by compound interest. We *could* live in a world where asset allocations were ordinarily given as a mix of stocks, bonds, real estate, and Berkshire Hathaway. We don't live in that world for a number of reasons: financial advisors not wanting to make themselves appear irrelevant, strange personal preferences on the part of Buffett . . .

The economy doesn't always do the obvious thing, like flow money into Buffett until his returns approach the average return of the economy. Interest rate differences much higher than 0.5%, on matters that people care about far more intensely than Science, are ignored if they're not presented in exactly the right format to be seized.

And it's not easy for individual scientists or groups to capture the value created by scientific progress. Did Einstein die with 0.1% of the value that he created? Engelbart in particular doesn't seem to have *tried* to be Bill Gates, at least not as far as I know.

With that in mind—in one sense Engelbart succeeded at a good portion of what he *actually set out* to do: computer mice *did* take over the world.

But it was a broad slow cascade that mixed into the usual exponent of economic growth. Not a concentrated fast FOOM. To produce a concentrated FOOM, you've got to be able to swallow as much as possible of the processes *driving* the FOOM *into* the FOOM. Otherwise you can't improve those processes and you can't cascade through them and your k goes down. Then your interest rates won't even be

as much higher than normal as, say, Warren Buffett's. And there's no grail to be *won*, only profits to be made: If you have no realistic hope of beating the world, you may as well join it.

* * *

Eliezer Yudkowsky

Humanity is in a FOOM relative to the rest of the biosphere but of course it doesn't seem ridiculously fast to *us*; the question from our standpoint is whether a brain in a box in a basement can go FOOM relative to human society. Anyone who thinks that, because we're already growing at a high rate, the distinction between that and a nanotech-capable superintelligence must not be very important is being just a little silly. It may not even be wise to call them by the same name, if it tempts you to such folly—and so I would suggest reserving "FOOM" for things that go very fast relative to *you*.

For the record, I've been a coder and judged myself a reasonable hacker—set out to design my own programming language at one point, which I say not as a mark of virtue but just to demonstrate that I was in the game. (Gave it up when I realized AI wasn't about programming languages.)

See original post for all comments.

* * *

1. Engelbart, *Augmenting Human Intellect*.

2. Moishe Lettvin, "The Windows Shutdown Crapfest," *Moishe's Blog* (blog), November 24, 2006, http://moishelettvin.blogspot.com/2006/11/windows-shutdown-crapfest.html.

3. K. Eric Drexler, *Engines of Creation* (Garden City, NY: Anchor, 1986).

25

Total Nano Domination

Eliezer Yudkowsky

27 November 2008

Followup to: Engelbart: Insufficiently Recursive

The computer revolution had cascades and insights aplenty. Computer tools are routinely used to create tools, from using a C compiler to write a Python interpreter to using theorem-proving software to help design computer chips. I would not *yet* rate computers as being very deeply *recursive*—I don't think they've improved our own thinking processes even so much as the Scientific Revolution—*yet*. But some of the ways that computers are used to improve computers verge on being repeatable (cyclic).

Yet no individual, no localized group, nor even country, managed to get a sustained advantage in computing power, compound the interest on cascades, and take over the world. There was never

a Manhattan moment when a computing advantage *temporarily* gave one country a supreme military advantage, like the US and its atomic bombs for that brief instant at the end of WW2. In computing there was no equivalent of "We've just crossed the sharp threshold of criticality, and now our pile doubles its neutron output every *two minutes*, so we can produce lots of plutonium and you can't."

Will the development of nanotechnology go the same way as computers—a smooth, steady developmental curve spread across many countries, no one project taking into itself a substantial fraction of the world's whole progress? Will it be more like the Manhattan Project, one country gaining a (temporary?) huge advantage at huge cost? Or could a small group with an initial advantage cascade and outrun the world?

Just to make it clear why we might worry about this for nanotech, rather than say car manufacturing—if you can build things from atoms, then the environment contains an unlimited supply of perfectly machined spare parts. If your molecular factory can build solar cells, it can acquire energy as well.

So full-fledged Drexlerian molecular nanotechnology (Wikipedia) can plausibly automate away much of the *manufacturing* in its *material* supply chain. If you already have nanotech, you may not need to consult the outside economy for inputs of energy or raw material.

This makes it more plausible that a nanotech group could localize off, and do its own compound interest away from the global economy. If you're Douglas Engelbart building better software, you still need to consult Intel for the hardware that runs your software, and the electric company for the electricity that powers your hardware. It would be

a *considerable expense* to build your own fab lab for your chips (that makes chips as good as Intel) and your own power station for electricity (that supplies electricity as cheaply as the utility company).

It's not just that this tends to entangle you with the fortunes of your trade partners, but also that—as an *UberTool Corp* keeping your trade secrets in-house—you can't improve the hardware you get, or drive down the cost of electricity, as long as these things are done outside. Your cascades can only go through what you do locally, so the more you do locally, the more likely you are to get a compound interest advantage. (Mind you, I don't think Engelbart could have gone FOOM even if he'd made his chips locally and supplied himself with electrical power—I just don't think the compound advantage on using computers to make computers is powerful enough to sustain $k > 1$.)

In general, the more capabilities are localized into one place, the less people will depend on their trade partners, the more they can cascade locally (apply their improvements to yield further improvements), and the more a "critical cascade"/FOOM sounds plausible.

Yet self-replicating nanotech is a very *advanced* capability. You don't get it right off the bat. Sure, lots of biological stuff has this capability, but this is a misleading coincidence—it's not that self-replication is *easy*, but that evolution, *for its own alien reasons*, tends to build it into everything. (Even individual cells, which is ridiculous.)

In the *run-up* to nanotechnology, it seems not implausible to suppose a continuation of the modern world. Today, many different labs work on small pieces of nanotechnology—fortunes entangled with their trade partners, and much of their research velocity coming from advances in other laboratories. Current nanotech labs are dependent on the outside world for computers, equipment, science, electricity,

and food; any single lab works on a small fraction of the puzzle, and contributes small fractions of the progress.

In short, so far nanotech is going just the same way as computing.

But it is a tad premature—I would even say that it crosses the line into the "silly" species of futurism—to exhale a sigh of relief and say, "Ah, that settles it—no need to consider any further."

We all know how exponential multiplication works: 1 microscopic nanofactory, 2 microscopic nanofactories, 4 microscopic nanofactories . . . let's say there's a hundred different groups working on self-replicating nanotechnology and one of those groups succeeds one week earlier than the others. Rob Freitas has calculated that some species of replibots could spread through the Earth in two days (even given what seem to me like highly conservative assumptions in a context where conservatism is not appropriate).[1]

So, even if the race seems very tight, whichever group gets replibots *first* can take over the world given a mere week's lead time—

Yet wait! Just having replibots doesn't let you take over the world. You need fusion weapons, or surveillance bacteria, or some other way to actually *govern*. That's a lot of matterware—a lot of design and engineering work. A replibot advantage doesn't equate to a weapons advantage, unless, somehow, the planetary economy has already published the open-source details of fully debugged weapons that you can build with your newfound private replibots. Otherwise, a lead time of one week might not be anywhere near enough.

Even more importantly—"self-replication" is not a binary, 0-or-1 attribute. Things can be partially self-replicating. You can have something that manufactures 25% of itself, 50% of itself, 90% of itself, or 99% of itself—but still needs one last expensive computer chip to

complete the set. So if you have twenty-five countries racing, sharing some of their results and withholding others, there isn't *one morning* where you wake up and find that one country has self-replication.

Bots become successively easier to manufacture; the factories get successively cheaper. By the time one country has bots that manufacture themselves from environmental materials, many other countries have bots that manufacture themselves from feedstock. By the time one country has bots that manufacture themselves entirely from feedstock, other countries have produced some bots using assembly lines. The nations also have all their old conventional arsenal, such as intercontinental missiles tipped with thermonuclear weapons, and these have deterrent effects against crude nanotechnology. No one ever gets a *discontinuous* military advantage, and the world is safe (?).

At this point, I do feel obliged to recall the notion of "burdensome details," that we're spinning a story out of many conjunctive details, any one of which could go wrong. This is not an argument in favor of anything in particular, just a reminder not to be seduced by stories that are too specific. When I contemplate the sheer raw power of nanotechnology, I don't feel confident that the fabric of society can even survive the *sufficiently plausible prospect* of its near-term arrival. If your intelligence estimate says that Russia (the new belligerent Russia under Putin) is going to get self-replicating nanotechnology in a year, what does that do to Mutual Assured Destruction? What if Russia makes a similar intelligence assessment of the US? What happens to the capital markets? I can't even foresee how our world will react to the *prospect* of various nanotechnological capabilities as they promise to be developed in the future's near future. Let alone envision how so-

ciety would *actually change* as full-fledged molecular nanotechnology was developed, even if it were developed gradually . . .

. . . but I suppose the Victorians might say the same thing about nuclear weapons or computers, and yet we still have a global economy—one that's actually lot more interdependent than theirs, thanks to nuclear weapons making small wars less attractive, and computers helping to coordinate trade.

I'm willing to believe in the possibility of a smooth, gradual ascent to nanotechnology, so that no one state—let alone any corporation or small group—ever gets a discontinuous advantage.

The main reason I'm willing to believe this is because of the difficulties of *design* and *engineering*, even after all manufacturing is solved. When I read Drexler's *Nanosystems*, I thought: "Drexler uses properly conservative assumptions everywhere I can see, except in one place—debugging. He assumes that any failed component fails visibly, immediately, and without side effects; *this* is not conservative."

In *principle*, we have complete control of our computers—every bit and byte is under human command—and yet it still takes an immense amount of engineering work on top of that to make the bits do what we want. This, and not any difficulties of manufacturing things once they *are* designed, is what takes an international supply chain of millions of programmers.

But we're *still* not out of the woods.

Suppose that, by a providentially incremental and distributed process, we arrive at a world of full-scale molecular nanotechnology—a world where *designs*, if not finished material goods, are traded among parties. In a global economy large enough

that no one actor, or even any one state, is doing more than a fraction of the total engineering.

It would be a *very* different world, I expect; and it's possible that my essay may have already degenerated into nonsense. But even if we still have a global economy after getting this far—then we're *still* not out of the woods.

Remember those ems? The emulated humans-on-a-chip? The uploads?

Suppose that, with molecular nanotechnology already in place, there's an international race for reliable uploading—with some results shared, and some results private—with many state and some nonstate actors.

And suppose the race is so tight that the first state to develop working researchers-on-a-chip only has a *one-day* lead time over the other actors.

That is—one day before anyone else, they develop uploads sufficiently undamaged, or capable of sufficient recovery, that the ems can carry out research and development. In the domain of, say, uploading.

There are other teams working on the problem, but their uploads are still a little off, suffering seizures and having memory faults and generally having their cognition degraded to the point of not being able to contribute. (Note: I think this whole future is a wrong turn and we should stay away from it; I am not endorsing this.)

But this one team, though—their uploads still have a few problems, but they're at least sane enough and smart enough to start . . . fixing their problems themselves?

If there's already full-scale nanotechnology around when this happens, then even with some inefficiency built in, the first uploads may be running at ten thousand times human speed. Nanocomputers are powerful stuff.

And in an hour, or around a year of internal time, the ems may be able to upgrade themselves to a hundred thousand times human speed and fix some of the remaining problems.

And in another hour, or ten years of internal time, the ems may be able to get the factor up to a million times human speed, and start working on intelligence enhancement . . .

One could, of course, voluntarily publish the improved-upload protocols to the world and give everyone else a chance to join in. But you'd have to trust that not a single one of your partners were holding back a trick that lets them run uploads at ten times your own maximum speed (once the bugs were out of the process). That kind of advantage could snowball quite a lot, in the first sidereal day.

Now, if uploads are *gradually* developed *at a time when computers are too slow to run them quickly*—meaning, *before* molecular nanotech and nanofactories come along—then this whole scenario is averted; the first high-fidelity uploads, running at a hundredth of human speed, will grant no special advantage. (Assuming that no one is pulling any spectacular snowballing tricks with intelligence enhancement—but they would have to snowball fast and hard to confer advantage on a small group running at low speeds. The same could be said of brain-computer interfaces, developed before or after nanotechnology, if running in a small group at merely human speeds. I would credit their world takeover, but I suspect Robin Hanson wouldn't at this point.)

Now, I don't *really* believe in any of this—this whole scenario, this whole world I'm depicting. In real life, I'd expect someone to brute-force an unFriendly AI on one of those super-ultimate-nanocomputers, followed in short order by the end of the world. But that's a separate issue. And this whole world seems too much like our own, after too much technological change, to be realistic to me. World government with an insuperable advantage? Ubiquitous surveillance? I don't like the ideas, but both of them would change the game dramatically . . .

But the real point of this essay is to illustrate a point more important than nanotechnology: **as optimizers become more self-swallowing, races between them are more unstable.**

If you sent a modern computer back in time to 1950—containing many modern software tools in compiled form, but no future history or declaratively stored future science—I would guess that the recipient could *not* use it to take over the world. Even if the USSR got it. Our computing *industry* is a very powerful thing, but it relies on a supply chain of chip factories.

If someone got a future *nanofactory* with a library of future nanotech applications—including designs for things like fusion power generators and surveillance bacteria—they might really be able to *take over the world*. The nanofactory swallows its own supply chain; it incorporates replication within itself. If the owner fails, it won't be for lack of factories. It will be for lack of ability to develop new matterware fast enough, and apply existing matterware fast enough, to take over the world.

I'm not saying that nanotech *will* appear from nowhere with a library of designs—just making a point about concentrated power and the instability it implies.

Think of all the tech news that you hear about once—say, an article on *Slashdot* about yada yada 50% improved battery technology—and then you never hear about again, because it was too expensive or too difficult to manufacture.

Now imagine a world where the news of a 50% improved battery technology comes down the wire, and the head of some country's defense agency is sitting down across from engineers and intelligence officers and saying, "We have five minutes before all of our rival's weapons are adapted to incorporate this new technology; how does that affect our balance of power?" Imagine that happening as often as "amazing breakthrough" articles appear on *Slashdot*.

I don't mean to doomsay—the Victorians would probably be pretty surprised we haven't blown up the world with our ten-minute ICBMs, but we don't live in their world—well, maybe doomsay just a little—but the point is: *It's less stable*. Improvements cascade faster once you've swallowed your manufacturing supply chain.

And if you sent back in time a single nanofactory, *and* a single upload living inside it—then the world might end in five minutes or so, as we bios measure time.

The point being not that an upload *will* suddenly appear, but that now you've swallowed your supply chain *and* your R&D chain.

And so this world is correspondingly more unstable, even if all the actors start out in roughly the same place. Suppose a state manages to get one of those *Slashdot*-like technology improvements—only this one lets uploads think 50% faster—and they get it fifty minutes before

anyone else, at a point where uploads are running ten thousand times as fast as human (50 mins. ≈1 year)—and in that extra half year, the uploads manage to find another couple of 50% improvements . . .

Now, you *can* suppose that all the actors are all trading all of their advantages and holding nothing back, so everyone stays nicely synchronized.

Or you can suppose that enough trading is going on that most of the research any group benefits from comes from *outside* that group, and so a 50% advantage for a local group doesn't cascade much.

But again, that's not the point. The point is that in modern times, with the modern computing industry, where commercializing an advance requires building a new computer factory, a bright idea that has gotten as far as showing a 50% improvement in the laboratory is merely one more article on *Slashdot*.

If everything could instantly be rebuilt via nanotech, that laboratory demonstration could precipitate an instant international military crisis.

And if there are uploads around, so that a cute little 50% advancement in a certain kind of hardware recurses back to imply *50% greater speed at all future research*—then this *Slashdot* article could become the key to world domination.

As systems get more self-swallowing, they cascade harder; and even if all actors start out equivalent, races between them get much more unstable. I'm not claiming it's impossible for that world to be stable. The Victorians might have thought that about ICBMs. But that subjunctive world contains *additional* instability compared to our own and would need *additional* centripetal forces to end up as stable as our own.

I expect Robin to disagree with some part of this essay, but I'm not sure which part or how.

* * *

Robin Hanson

Well, at long last you finally seem to be laying out the heart of your argument. Dare I hope that we can conclude our discussion by focusing on these issues, or are there yet more layers to this onion?

Eliezer Yudkowsky

It takes two people to make a disagreement; I don't *know* what the heart of my argument is from your perspective!

This essay treats the simpler and less worrisome case of nanotech. Quickie preview of AI:

- When you upgrade to AI there are harder faster cascades because the development idiom is even more recursive, and there is an overhang of hardware capability we don't understand how to use.
- There are probably larger development gaps between projects due to a larger role for insights.
- There are more barriers to trade between AIs, because of the differences of cognitive architecture—different AGI projects have far less in common today than nanotech projects, and there is very little sharing of cognitive content even in ordinary AI.
- Even if AIs trade improvements among themselves, there's a huge barrier to applying those improvements to human brains, uncrossable short of very advanced technology for uploading and extreme upgrading.

- So even if many unFriendly AI projects are developmentally synchronized and mutually trading, they may come to their own compromise, do a synchronized takeoff, and eat the biosphere; without caring for humanity, humane values, or any sort of existence for themselves that we regard as worthwhile . . .

But I don't know if you regard any of that as the *important* part of the argument, or if the key issue in our disagreement happens to be already displayed *here*. If it's here, we should resolve it here, because nanotech is much easier to understand.

Robin Hanson

In your one upload team a day ahead scenario, by "full-scale nanotech" you apparently mean oriented around very local production. That is, they don't suffer much efficiency reduction by building everything themselves on-site via completely automated production. The overall efficiency of this tech with available cheap feedstocks allows a doubling time of much less than one day. And in much less than a day this tech plus feedstocks cheaply available to this one team allow it to create more upload equivalents (scaled by speedups) than all the other teams put together. Do I understand you right?

Eliezer Yudkowsky

As I understand nanocomputers, it shouldn't really take all that *much* nanocomputer material to run more uploads than a bunch of bios—like, a cubic meter of nanocomputers total, and a megawatt of electricity, or something like that. The key point is that you have such-and-such amount of nanocomputers available—it's not a focus on material production per se.

Also, bear in mind that I already acknowledged that you could have a slow run-up to uploading such that there's no hardware overhang when the very first

uploads capable of doing their own research are developed—the one-day lead and the fifty-minute lead are two different scenarios above.

See original post for all comments.

* * *

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

26

Dreams of Autarky

Robin Hanson

27 November 2008

Selections from my 1999 essay "Dreams of Autarky":[1]

> [Here is] an important common bias on "our" side, i.e., among those who expect specific very large changes. . . . Futurists tend to expect an unrealistic degree of autarky, or independence, within future technological and social systems. The cells in our bodies are largely-autonomous devices and manufacturing plants, producing most of what they need internally. . . . Small tribes themselves were quite autonomous. . . . Most people are not very aware of, and so have not fully to terms with their new inter-dependence. For example, people are surprisingly willing to restrict trade between nations, not realizing how much their wealth depends on such trade. . . . Futurists commonly neglect this interdependence . . . they picture their future political and

economic unit to be the largely self-sufficient small tribe of our evolutionary heritage. . . . [Here are] some examples. . . .

[Many] imagine space economies almost entirely self-sufficient in mass and energy. . . . It would be easier to create self-sufficient colonies under the sea, or in Antarctica, yet there seems to be little prospect of or interest in doing so anytime soon. . . .

Eric Drexler . . . imagines manufacturing plants that are far more independent than in our familiar economy. . . . To achieve this we need not just . . . control of matter at the atomic level, but also the *complete* automation of the manufacturing process, all embodied in a single device . . . complete with quality control, waste management, and error recovery. This requires "artificial intelligence" far more advanced than we presently possess. . . .

Knowledge is [now] embodied in human-created software and hardware, and in human workers trained for specific tasks. . . . It has usually been cheaper to leave the CPU and communication intensive tasks to machines, and leave the tasks requiring general knowledge to people.

Turing-test artificial intelligence instead imagines a future with many large human-created software modules . . . far more independent, i.e., less dependent on context, than existing human-created software. . . .

[Today] innovations and advances in each part of the world [depends] on advances made in all other parts of the world. . . . Visions of a local singularity, in contrast, imagine that sudden technological advances in one small group essentially allow that group to suddenly grow big enough to take over everything. . . . The key common assumption is that of a very powerful but autonomous area of technology. Overall progress in that area must depend only on advances in this area, advances that a small group of researchers can

continue to produce at will. And great progress in this area alone must be sufficient to let a small group essentially take over the world. . . .

[Crypto credential] dreams imagine that many of our relationships will be exclusively digital, and that we can keep these relations independent by separating our identity into relationship-specific identities. . . . It is hard to imagine potential employers not asking to know more about you, however. . . . Any small information leak can be enough to allow others to connect your different identities. . . .

[Consider also] complaints about the great specialization in modern academic and intellectual life. People complain that ordinary folks should know more science, so they can judge simple science arguments for themselves. . . . Many want policy debates to focus on intrinsic merits, rather than on appeals to authority. Many people wish students would study a wider range of subjects, and so be better able to see the big picture. And they wish researchers weren't so penalized for working between disciplines, or for failing to cite every last paper someone might think is related somehow.

It seems to me plausible to attribute all of these dreams of autarky to people not yet coming fully to terms with our newly heightened interdependence. . . . We picture our ideal political unit and future home to be the largely self-sufficient small tribe of our evolutionary heritage. . . . I suspect that future software, manufacturing plants, and colonies will typically be much more dependent on everyone else than dreams of autonomy imagine. Yes, small isolated entities are getting more capable, but so are small non-isolated entities, and the latter remain far more capable than the former. The riches that come from a worldwide division of labor have rightly seduced us away from many of

> our dreams of autarky. We may fantasize about dropping out of the rat race and living a life of ease on some tropical island. But very few of us ever do.
>
> So academic specialists may dominate intellectual progress, and world culture may continue to overwhelm local variations. Private law and crypto-credentials may remain as marginalized as utopian communities have always been. Manufacturing plants may slowly get more efficient, precise, and automated without a sudden genie nanotech revolution. Nearby space may stay uncolonized until we can cheaply send lots of mass up there, while distant stars may remain uncolonized for a long long time. And software may slowly get smarter, and be collectively much smarter than people long before anyone bothers to make a single module that can pass a Turing test.

The relevance to my discussion with Eliezer should be obvious. My next post will speak more directly.

* * *

Eliezer Yudkowsky

We generally specialize when it comes to bugs in computer programs—rather than monitoring their behavior and fixing them ourselves, we inform the central development authority for that program of the problem and rely on them to fix it everywhere.

The benefit from automation depends on the amount of human labor already in the process, à la the bee-sting principle of poverty. Automating one operation while many others are still human-controlled is a marginal improvement, because you can't run at full speed or fire your human resources department until you've gotten rid of all the humans.

The incentive for automation depends on the number of operations being performed. If you're doing something a trillion times over, it has to be automatic. We pay whatever energy cost is required to make transistor operations on chips fully reliable, because it would be impossible to have a chip if each transistor required human monitoring. DNA sequencing is increasingly automated as we try to do more and more of it.

With nanotechnology it is more *possible* to automate because you are designing all the machine elements of the system on a finer grain, closer to the level of physical law where interactions are perfectly regular, and more importantly, closing the system: no humans wandering around on your manufacturing floor.

And the *incentive* to automate is tremendous because of the gigantic number of operations you want to perform, and the higher levels of organization you want to build on top—it is akin to the incentive to automate the internal workings of a computer chip.

Now with all that said, I find it extremely plausible that, as with DNA sequencing, we will only see an increasing degree of automation over time, rather than a sudden *fully* automated system appearing *ab initio*. The operators will be there, but they'll handle larger and larger systems, and finally, in at least some cases, they'll disappear. Not assembly line workers, sysadmins. Bugs will continue to be found but their handling will be centralized and one-off rather than local and continuous. The system will behave more like the inside of a computer chip than the inside of a factory.

—Such would be my guess, not to materialize instantly but as a trend over time.

Robin Hanson

Eliezer, yes, the degree of automation will probably increase incrementally. As I explore somewhat here,[2] there is also the related issue of the degree of local production, vs. importing inputs made elsewhere. A high degree of automation need not induce a high degree of local production. Perhaps each different

group specializes in automating certain aspects of production, and they coordinate by sending physical inputs to each other.

Eliezer Yudkowsky

Robin, numerous informational tasks can be performed far more quickly by special-purpose hardware, arguably analogous to more efficient special-purpose molecular manufacturers. The cost of shipping information is incredibly cheap. Yet the typical computer contains a CPU and a GPU and does not farm out hard computational tasks to distant specialized processors. Even when we do farm out some tasks, mostly for reason of centralizing information rather than computational difficulty, the tasks are still given to large systems of conventional CPUs. Even supercomputers are mostly made of conventional CPUs.

This proves nothing, of course; but it is worth observing of the computational economy, in case you have some point that differentiates it from the nanotech economy. Are you sure you're not being prejudiced by the sheer *traditionalness* of moving physical inputs around through specialized processors?

Robin Hanson

Eliezer, both computing and manufacturing are old enough now to be "traditional"; I expect each mode of operation is reasonably well adapted to current circumstances. Yes, future circumstances will change, but do we really know in which direction? Manufacturing systems may well also now ship material over distances "for reason of centralizing information."

See original post for all comments.

* * *

1. Robin Hanson, "Dreams of Autarky" (Unpublished manuscript, September 1999), last revised September 2001, http://hanson.gmu.edu/dreamautarky.html.

2. Robin Hanson, "Five Nanotech Social Scenarios," in *Nanotechnology: Societal Implications—Individual Perspectives*, ed. Mihail C. Roco and William Sims Bainbridge (Dordrecht, The Netherlands: Springer, 2007), 109–113.

27

Total Tech Wars

Robin Hanson

29 November 2008

Eliezer Thursday:

> Suppose . . . the first state to develop working researchers-on-a-chip, only has a *one-day* lead time. . . . If there's already full-scale nanotechnology around when this happens . . . in an hour . . . the ems may be able to upgrade themselves to a hundred thousand times human speed, . . . and in another hour . . . get the factor up to a million times human speed, and start working on intelligence enhancement. . . . One could, of course, voluntarily publish the improved-upload protocols to the world and give everyone else a chance to join in. But you'd have to trust that not a single one of your partners were holding back a trick that lets them run uploads at ten times your own maximum speed.

Carl Shulman Saturday and Monday:

> I very much doubt that any U.S. or Chinese President who understood the issues would fail to nationalize a for-profit firm under those circumstances. . . . It's also how a bunch of social democrats, or libertarians, or utilitarians, might run a project, knowing that a very likely alternative is the crack of a future dawn and burning the cosmic commons, with a lot of inequality in access to the future, and perhaps worse. Any state with a lead on bot development that can ensure the bot population is made up of nationalists or ideologues (who could monitor each other) could disarm the world's dictatorships, solve collective action problems. . . . [For] biological humans [to] retain their wealth as capital holders in his scenario, ems must be obedient and controllable enough. . . . But if such control is feasible, then a controlled em population being used to aggressively create a global singleton is also feasible.

***Every* new technology brings social disruption.** While new techs (broadly conceived) tend to increase the total pie, some folks gain more than others, and some even lose overall. The tech's inventors may gain intellectual property, it may fit better with some forms of capital than others, and those who first foresee its implications may profit from compatible investments. So any new tech can be framed as a conflict, between opponents in a race or war.

***Every* conflict can be framed as a total war.** If you believe the other side is totally committed to total victory, that surrender is unacceptable, and that all interactions are zero-sum, you may conclude your side must never cooperate with them, nor tolerate much internal dissent or luxury. All resources must be devoted to growing more resources and to fighting them in every possible way.

A total war is a self-fulfilling prophecy; a total war exists exactly when any substantial group believes it exists. And total wars need not be "hot." Sometimes your best war strategy is to grow internally, or wait for other forces to wear opponents down, and only at the end convert your resources into military power for a final blow.

These two views can be combined in ***total tech wars***. The pursuit of some particular tech can be framed as a crucial battle in our war with them; we must not share any of this tech with them, nor tolerate much internal conflict about how to proceed. We must race to get the tech first and retain dominance.

Tech transitions produce variance in who wins more. If you are ahead in a conflict, added variance reduces your chance of winning, but if you are behind, variance increases your chances. So the prospect of a tech transition gives hope to underdogs, and fear to overdogs. The bigger the tech, the bigger the hopes and fears.

In 1994 I said that, while our future vision usually fades into a vast fog of possibilities, brain emulation "excites me because it seems an exception to this general rule—more like a crack of dawn than a fog, like a sharp transition with sharp implications regardless of the night that went before."[1] In fact, brain emulation is the largest tech disruption I can foresee (as more likely than not to occur). So yes, one might frame brain emulation as a total tech war, bringing hope to some and fear to others.

And yes, the size of that disruption is uncertain. For example, an em transition could go relatively smoothly if scanning and cell modeling techs were good enough well before computers were cheap enough. In this case em workers would gradually displace human workers as computer costs fell. If, however, one group suddenly had

the last key modeling breakthrough when em computer costs were far below human wages, that group could gain enormous wealth, to use as they saw fit.

Yes, if such a winning group saw itself in a total war, it might refuse to cooperate with others and devote itself to translating its breakthrough into an overwhelming military advantage. And yes, if you had enough reason to think powerful others saw this as a total tech war, you might be forced to treat it that way yourself.

Tech transitions that create whole new populations of beings can also be framed as total wars between the new beings and everyone else. If you framed a new-being tech this way, you might want to prevent or delay its arrival, or try to make the new beings "friendly" slaves with no inclination or ability to war.

But note: this em tech has no intrinsic connection to a total war other than that it is a big transition whereby some could win big! Unless you claim that all big techs produce total wars, you need to say why this one is different.

Yes, you can frame big techs as total tech wars, but surely **it is far better that tech transitions *not be framed as total wars***. The vast majority of conflicts in our society take place within systems of peace and property, where local winners only rarely hurt others much by spending their gains. It would be far better if new em tech firms sought profits for their shareholders, and allowed themselves to become interdependent because they expected other firms to act similarly.

Yes, we must be open to evidence that other powerful groups will treat new techs as total wars. But **we must avoid *creating* a total war by sloppy discussion of it as a possibility**. We should not take others' discussions of this possibility as strong evidence that they will treat a

tech as total war, nor should we discuss a tech in ways that others could reasonably take as strong evidence we will treat it as total war. Please, "give peace a chance."

Finally, note our many biases to overtreat techs as wars. There is a vast graveyard of wasteful government projects created on the rationale that a certain region must win a certain tech race/war. Not only do governments do a lousy job of guessing which races they could win, they also overestimate both first mover advantages and the disadvantages when others dominate a tech. Furthermore, as I posted Wednesday:

> We seem primed to confidently see history as an inevitable march toward a theory-predicted global conflict with an alien united *them* determined to oppose our core symbolic values, making infeasible overly risky overconfident plans to oppose them.

* * *

Eliezer Yudkowsky

I generally refer to this scenario as "winner take all" and had planned a future post with that title.

I'd never have dreamed of calling it a "total tech war" because that sounds much too combative, a phrase that might spark violence even in the near term. It also doesn't sound accurate, because a winner-take-all scenario doesn't imply destructive combat or any sort of military conflict.

I moreover defy you to look over my writings and find any case where I ever used a phrase as inflammatory as "total tech war."

I think that, in this conversation and in the debate as you have just now framed it, "*Tu quoque!*" is actually justified here.

Anyway—as best as I can tell, the *natural* landscape of these technologies, *which introduces disruptions much larger than farming or the Internet*, is without special effort winner-take-all. It's not a question of ending up in that scenario by making special errors. We're just there. Getting out of it would imply special difficulty, not getting into it, and I'm not sure that's possible—such would be the stance I would try to support.

Also:

If you try to look at it from my perspective, then you can see that I've gone to *tremendous* lengths to defuse both the reality and the appearance of conflict between altruistic humans over which AI should be built. "Coherent Extrapolated Volition" is extremely meta; if all *competent and altruistic* Friendly AI projects think this meta, they are far more likely to find themselves able to cooperate than if one project says "Libertarianism!" and another says "Social democracy!"

On the other hand, the AGI projects run by the meddling dabblers *do* just say "Libertarianism!" or "Social democracy!" or whatever strikes their founder's fancy. And so far as I can tell, as a *matter of simple fact*, an AI project run at that level of competence will destroy the world. (It wouldn't be a good idea even if it worked as intended, but that's a separate issue.)

As a matter of simple decision theory, it seems to me that an unFriendly AI which has just acquired a decisive first-mover advantage is faced with the following payoff matrix:

Share Tech, Trade	10 utilons
Take Over Universe	1,000 utilons

As a matter of simple decision theory, I expect an unFriendly AI to take the second option.

Do you agree that *if* an unFriendly AI gets nanotech and no one else has nanotech, it will take over the world rather than trade with it?

Or is this statement something that is true but forbidden to speak?

Eliezer Yudkowsky

We could be in any of the three following domains:

1. The tech landscape is naturally smooth enough that, even if participants don't share technology, there is no winner take all.

2. The tech landscape is somewhat steep. If participants don't share technology, one participant will pull ahead and dominate all others via compound interest. If they share technology, the foremost participant will only control a small fraction of the progress and will not be able to dominate all other participants.

3. The tech landscape contains upward cliffs, and/or progress is naturally hard to share. Even if participants make efforts to trade progress up to time T, one participant will, after making an additional discovery at time $T+1$, be faced with at least the *option* of taking over the world. Or it is plausible for a single participant to withdraw from the trade compact, and either (a) accumulate private advantages while monitoring open progress or (b) do its own research, and still take over the world.

(Two) is the only regime where you can have self-fulfilling prophecies. I think nanotech is probably in (2) but contend that AI lies naturally in (3).

Robin Hanson

Eliezer, if everything is at stake then "winner take all" is "total war"; it doesn't really matter if they shoot you or just starve you to death. The whole point of this post is to note that anything can be seen as "winner-take-all" just by expecting others to see it that way. So if you want to say that a particular tech is *more* winner-take-all than usual, you need an argument based on more than just this effect. And if you want to argue it is *far* more so than any other tech humans have ever seen, you need a damn good additional argument. It is possible that you could make such an argument work based on the "tech landscape" considerations you mention, but I haven't seen that yet. So consider this

post to be yet another reminder that I await hearing your core argument; until then I set the stage with posts like this.

To answer your direct questions, I am not suggesting forbidding speaking of anything, and if "unfriendly AI" is *defined* as an AI who sees itself in a total war, then sure, it would take a total war strategy of fighting not trading. But you haven't actually defined "unfriendly" yet. . . .

See original post for all comments.

* * *

1. Hanson, "If Uploads Come First."

28

Singletons Rule OK

Eliezer Yudkowsky

30 November 2008

Reply to: Total Tech Wars

How *does* one end up with a persistent disagreement between two rationalist-wannabes who are both aware of Aumann's Agreement Theorem and its implications?

Such a case is likely to turn around two axes: object-level incredulity ("no matter *what* AAT says, proposition X can't *really* be true") and meta-level distrust ("they're trying to be rational despite their emotional commitment, but are they really capable of that?").

So far, Robin and I have focused on the object level in trying to hash out our disagreement. Technically, I can't speak for Robin; but at least in my *own* case, I've acted thus because I anticipate that a meta-level argument about trustworthiness wouldn't lead anywhere

interesting. Behind the scenes, I'm doing what I can to make sure my brain is actually capable of updating, and presumably Robin is doing the same.

(The linchpin of my own current effort in this area is to tell myself that I ought to be learning something while having this conversation, and that I shouldn't miss any scrap of original thought in it—the Incremental Update technique. Because I can genuinely believe that a conversation like this should produce new thoughts, I can turn that feeling into genuine attentiveness.)

Yesterday, Robin inveighed hard against what he called "total tech wars," and what I call "winner-take-all" scenarios:

> If you believe the other side is totally committed to total victory, that surrender is unacceptable, and that all interactions are zero-sum, you may conclude your side must never cooperate with them, nor tolerate much internal dissent or luxury.

Robin and I both have emotional commitments and we both acknowledge the danger of that. There's nothing irrational about feeling, *per se*; only *failure to update* is blameworthy. But Robin seems to be *very* strongly against winner-take-all technological scenarios, and I don't understand why.

Among other things, I would like to ask if Robin has a Line of Retreat set up here—if, regardless of how he estimates the *probabilities*, he can *visualize what he would do* if a winner-take-all scenario were true.

Yesterday Robin wrote:

> Eliezer, if everything is at stake then "winner take all" is "total war"; it doesn't really matter if they shoot you or just starve you to death.

We both have our emotional commitments, but I don't quite understand this reaction.

First, to me it's obvious that a "winner-take-all" *technology* should be defined as one in which, *ceteris paribus*, a local entity tends to end up with the *option* of becoming one kind of Bostromian singleton—the decision maker of a global order in which there is a single decision-making entity at the highest level.[1] (A superintelligence with unshared nanotech would count as a singleton; a federated world government with its own military would be a different kind of singleton; or you can imagine something like a galactic operating system with a root account controllable by 80% majority vote of the populace, *et cetera*.)

The winner-take-all *option* is created by properties of the technology landscape, which is not a moral stance. Nothing is said about an agent with that *option actually* becoming a singleton. Nor about *using* that power to shoot people, or reuse their atoms for something else, or grab all resources and let them starve (though "all resources" should include their atoms anyway).

Nothing is yet said about various patches that could try to avert a *technological* scenario that contains upward cliffs of progress—e.g., binding agreements enforced by source code examination or continuous monitoring—in advance of the event. (Or if you think that rational agents cooperate on the Prisoner's Dilemma, so much work might not be required to coordinate.)

Superintelligent agents *not* in a humanish moral reference frame—AIs that are just maximizing paperclips or sorting pebbles—who happen on the option of becoming a Bostromian Singleton, and who have *not* previously executed any somehow-binding treaty, will *ceteris paribus* choose to grab all resources in service of their utility function, including the atoms now comprising humanity. I don't see how you could reasonably deny this! It's a straightforward decision-theoretic choice between payoff 10 and payoff 1,000!

But conversely, there are possible agents in mind-design space who, given the *option* of becoming a singleton, will *not* kill you, starve you, reprogram you, tell you how to live your life, or even meddle in your destiny unseen. See Bostrom's (short) paper on the possibility of good and bad singletons of various types.[2]

If Robin thinks it's *impossible* to have a Friendly AI or maybe even any sort of benevolent superintelligence at all, even the descendants of human uploads—if Robin is assuming that superintelligent agents *will* act according to roughly selfish motives, and that *only* economies of trade are necessary and sufficient to prevent holocaust—then Robin may have no Line of Retreat open as I try to argue that AI has an upward cliff built in.

And in this case, it might be time well spent to first address the question of whether Friendly AI is a reasonable thing to try to accomplish, so as to create that line of retreat. Robin and I are both trying hard to be rational despite emotional commitments; but there's no particular reason to *needlessly* place oneself in the position of trying to persuade, or trying to accept, that everything of value in the universe is certainly doomed.

For me, it's particularly hard to understand Robin's position in this, because for me the *non*-singleton future is the one that is obviously abhorrent.

If you have lots of entities with root permissions on matter, any of whom has the physical capability to attack any other, then you have entities spending huge amounts of precious negentropy on defense and deterrence. If there's no centralized system of property rights in place for selling off the universe to the highest bidder, then you have a race to burn the cosmic commons,[3] and the degeneration of the vast majority of all agents into rapacious hardscrapple frontier replicators.[4]

To me this is a vision of *futility*—one in which a future light cone that *could* have been full of happy, safe agents having complex fun is mostly wasted by agents trying to seize resources and defend them so they can send out seeds to seize more resources.

And it should also be mentioned that any future in which slavery or child abuse is *successfully* prohibited is a world that has *some* way of preventing agents from doing certain things with their computing power. There are vastly worse possibilities than slavery or child abuse opened up by future technologies, which I flinch from referring to even as much as I did in the previous sentence. There are things I don't want to happen to *anyone*—including a population of a septillion captive minds running on a star-powered matrioshka brain that is owned, and *defended* against all rescuers, by the mind-descendant of Lawrence Bittaker (serial killer, a.k.a. "Pliers"). I want to *win* against the horrors that exist in this world and the horrors that could exist in tomorrow's world—to have them never happen ever

again, or, for the *really* awful stuff, never happen in the first place. And that victory requires the Future to have certain *global* properties.

But there are other ways to get singletons besides falling up a technological cliff. So that would be my Line of Retreat: If minds can't self-improve quickly enough to take over, then try for the path of uploads setting up a centralized Constitutional operating system with a root account controlled by majority vote, or something like that, to prevent their descendants from *having* to burn the cosmic commons.

So for me, *any satisfactory outcome* seems to necessarily involve, if not a singleton, the existence of certain stable *global* properties upon the future—sufficient to *prevent* burning the cosmic commons, *prevent* life's degeneration into rapacious hardscrabble frontier replication, and *prevent* supersadists torturing septillions of helpless dolls in private, obscure star systems.

Robin has written about burning the cosmic commons and rapacious hardscrapple frontier existences. This doesn't imply that Robin approves of these outcomes. But Robin's strong rejection even of winner-take-all *language* and *concepts* seems to suggest that our emotional commitments are something like 180 degrees opposed. Robin seems to feel the same way about singletons as I feel about singletons.

But *why*? I don't think our real values are that strongly opposed—though we may have verbally described and attention-prioritized those values in different ways.

* * *

James Miller

You and Robin seem to be focused on different time periods. Robin is claiming that after ems are created one group probably won't get a dominant position. You are saying that post-intelligence-explosion (or at least post one day before the intelligence explosion) there will be either one dominant group or a high likelihood of total war. You are not in conflict if there is a large time gap between when we first have ems and when there is a intelligence explosion.

I wrote in this post that such a gap is likely: Billion Dollar Bots.

Robin Hanson

Eliezer, sometimes in a conversation one needs a rapid back and forth, often to clarify what exactly people mean by things they say. In such a situation a format like the one we are using, long daily blog posts, can work particularly badly. In my last post I was trying in part to get you to become clearer about what you meant by what you now call a "winner-take-all" tech, especially to place it on a continuum with other familiar techs. (And once we are clear on what it means, then I want arguments suggesting that an AI transition would be such a thing.) I suggested talking about outcome variance induced by a transition. If you now want to use that phrase to denote "a local entity tends to end up with the option of becoming one kind of Bostromian singleton," then we need new terms to refer to the "properties of the technology landscape" that might lead to such an option.

I am certainly not assuming it is impossible to be "friendly" though I can't be sure without knowing better what that means. I agree that it is not obvious that we would not want a singleton, if we could choose the sort we wanted. But I am, as you note, quite wary of the sort of total war that might be required to create a singleton. But before we can choose among options we need to get clearer on what the options are. . . .

Robin Hanson

Oh, to answer Eliezer's direct question directly, if I know that I am in a total war, I fight. I fight to make myself, or if that is impossible those who most share my values, win.

Eliezer Yudkowsky

> Sometimes in a conversation one needs a rapid back and forth . . .

Yeah, unfortunately I'm sort of in the middle of resetting my sleep cycle at the moment so I'm out of sync with you for purposes of conducting rapid-fire comments. Should be fixed in a few days. . . .

There are clear differences of worldview clashing here, which have nothing to do with the speed of an AI takeoff per se, but rather have something to do with what kind of technological progress parameters imply what sort of consequences. I was talking about large localized jumps in capability; you made a leap to total war. I can guess at some of your beliefs behind this but it would only be a guess. . . .

> Oh, to answer Eliezer's direct question directly, if I know that I am in a total war, I fight. I fight to make myself, or if that is impossible those who most share my values, win.

That's not much of a Line of Retreat. It would be like my saying, "Well, if a hard takeoff is impossible, I guess I'll try to make sure we have as much fun as we can in our short lives." If I *actually* believed an AI hard takeoff were impossible, I wouldn't pass directly to the worst-case scenario and give up on all other hopes. I would pursue the path of human intelligence enhancement, or uploading, or nontakeoff AI, and promote cryonics more heavily.

If you *actually* came to believe in large localized capability jumps, I do *not* think you would say, "Oh, well, guess I'm inevitably in a total war, now I need to fight a zero-sum game and damage all who are not my allies as much as possible." I think you would say, "Okay, so, how do we *avoid* a total war in this

kind of situation?" If you can work out in advance what you would do then, *that's* your line of retreat.

I'm sorry for this metaphor, but it just seems like a very useful and standard one if one can strip away the connotations: suppose I asked a theist to set up a Line of Retreat if there is no God, and they replied, "Then I'll just go through my existence trying to ignore the gaping existential void in my heart." That's not a line of retreat—that's a reinvocation of the same forces holding the original belief in place. I have the same problem with my asking, "Can you set up a line of retreat for yourself if there is a large localized capability jump?" and your replying, "Then I guess I would do my best to win the total war."

If you can make the implication *explicit*, and really look for loopholes, and fail to find them, then there is no line of retreat; but to me, at least, it looks like a line of retreat really should exist here.

Eliezer Yudkowsky

PS: As the above was a long comment and Robin's time is limited: if he does not reply to every line, no one should take that as evidence that no good reply exists. We also don't want to create a motive for people to try to win conversations by exhaustion.

Still, I'd like to hear a better line of retreat, even if it's one line like, I don't know, "Then I'd advocate regulations to slow down AI in favor of human enhancement" or something. Not that I'm saying this is a good idea, just something, anything, to break the link between AI hard takeoff and total moral catastrophe.

Robin Hanson

Eliezer, I'm very sorry if my language offends. If you tell the world you are building an AI and plan that post-foom it will take over the world, well, then that sounds to me like a declaration of total war on the rest of the world. Now

you might reasonably seek as large a coalition as possible to join you in your effort, and you might plan for the AI to not prefer you or your coalition in the acts it chooses. And you might reasonably see your hand as forced because other AI projects exist that would take over the world if you do not. But still, that take over the world step sure sounds like total war to me.

Oh, and on your "line of retreat," I might well join your coalition, given these assumptions. I tried to be clear about that in my Stuck In Throat post as well.

Eliezer Yudkowsky

If you're fighting a total war, then at some point, somewhere along the line, you should *at least stab someone in the throat*. If you don't do even that much, it's very hard for me to see it as a total war.

You described a total war as follows:

> If you believe the other side is totally committed to total victory, that surrender is unacceptable, and that all interactions are zero-sum, you may conclude your side must never cooperate with them, nor tolerate much internal dissent or luxury. All resources must be devoted to growing more resources and to fighting them in every possible way.

How is writing my computer program declaring "total war" on the world? Do I believe that "the world" is totally committed to total victory over me? Do I believe that surrender to "the world" is unacceptable—well, yes, I do. Do I believe that all interactions with "the world" are zero-sum? *Hell* no. Do I believe that I should never cooperate with "the world"? I do that every time I shop at a supermarket. Not tolerate internal dissent or luxury—both internal dissent and luxury sound good to me, I'll take both. All resources must be devoted to growing more resources and to fighting "the world" in every possible way? Mm . . . nah.

So you thus described a total war, and inveighed against it.

But then you applied the same term to the Friendly AI project, which has yet to stab a single person in the throat; and this, sir, I do not think is a fair description.

It is not a matter of indelicate language to be dealt with by substituting an appropriate euphemism. If I am to treat your words as consistently defined, then they are not, in this case, true.

Robin Hanson

Eliezer, I'm not very interested in arguing about which English words best describe the situation under consideration, at least if we are still unclear on the situation itself. Such words are just never that precise. Would you call a human stepping on an ant "total war," even if he wasn't trying very hard? From an aware ant's point of view it might seem total war, but perhaps you wouldn't say so if the human wasn't trying hard. But the key point is that the human could be in for a world of hurt if he displayed an intention to squash the ant and greatly underestimated the ant's ability to respond. So in a world where new AIs cannot in fact easily take over the world, AI projects that say they plan to have their AI take over the world could induce serious and harmful conflict.

See original post for all comments.

* * *

1. Nick Bostrom, "What is a Singleton?," *Linguistic and Philosophical Investigations* 5, no. 2 (2006): 48–54.

2. Ibid.

3. Robin Hanson, "Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Colonization" (Unpublished manuscript, July 1, 1998), accessed April 26, 2012, http://hanson.gmu.edu/filluniv.pdf.

4. Robin Hanson, "The Rapacious Hardscrapple Frontier," in *Year Million: Science at the Far Edge of Knowledge*, ed. Damien Broderick (New York: Atlas, 2008), 168–189, http://hanson.gmu.edu/hardscra.pdf.

29

Stuck In Throat

Robin Hanson

30 November 2008

Let me try again to summarize Eliezer's position, as I understand it, and what about it seems hard to swallow. I take Eliezer as saying:

> Sometime in the next few decades a human-level AI will probably be made by having a stupid AI make itself smarter. Such a process starts very slow and quiet, but eventually "fooms" very fast and then loud. It is likely to go from much stupider to much smarter than humans in less than a week. While stupid, it can be rather invisible to the world. Once smart, it can suddenly and without warning take over the world.
>
> The reason an AI can foom so much faster than its society is that an AI can change its basic mental architecture, and humans can't. How long any one AI takes to do this depends crucially on its initial architecture. Current architectures are so bad that an AI starting with them would take

> an eternity to foom. Success will come from hard math-like (and Bayes-net-like) thinking that produces deep insights giving much better architectures.
>
> A much smarter than human AI is basically impossible to contain or control; if it wants to it *will* take over the world, and then it *will* achieve whatever ends it has. One should have little confidence that one knows what those ends are from its behavior as a much less than human AI (e.g., as part of some evolutionary competition). Unless you have carefully proven that it wants what you think it wants, you have no idea what it wants.
>
> In such a situation, if one cannot prevent AI attempts by all others, then the only reasonable strategy is to try to be the first with a "friendly" AI, i.e., one where you really do know what it wants, and where what it wants is something carefully chosen to be as reasonable as possible.

I *don't* disagree with this last paragraph. But I do have trouble swallowing prior ones. The hardest to believe I think is that the AI will get smart so very rapidly, with a growth rate (e.g., doubling in an hour) so far out of proportion to prior growth rates, to what prior trends would suggest, and to what most other AI researchers I've talked to think. The key issues come from this timescale being so much shorter than team lead times and reaction times. This is the key point on which I await Eliezer's more detailed arguments.

Since I do accept that architectures can influence growth rates, I must also have trouble believing humans could find new AI architectures anytime soon that make this much difference. Some other doubts:

- Does a single "smarts" parameter really summarize most of the capability of diverse AIs?
- Could an AI's creators see what it wants by slowing down its growth as it approaches human level?
- Might faster brain emulations find it easier to track and manage an AI foom?

* * *

See original post for all comments.

30

Disappointment in the Future

Eliezer Yudkowsky
1 December 2008

This seems worth posting around now . . . As I've previously observed, futuristic visions are produced as entertainment, sold today and consumed today. A TV station interviewing an economic or diplomatic pundit doesn't bother to show what that pundit predicted three years ago and how the predictions turned out. Why would they? Futurism Isn't About Prediction.

But someone on the Longecity forum actually went and compiled a list of Ray Kurzweil's predictions in 1999 for the years 2000–2009.[1] We're not out of 2009 yet, but right now it's not looking good . . .

- Individuals primarily use portable computers.
- Portable computers have dramatically become lighter and thinner.

- Personal computers are available in a wide range of sizes and shapes, and are commonly embedded in clothing and jewelry, like wrist watches, rings, earrings and other body ornaments.
- Computers with a high-resolution visual interface range from rings and pins and credit cards up to the size of a thin book. People typically have at least a dozen computers on and around their bodies, which are networked using body LANs (local area networks).
- These computers monitor body functions, provide automated identity to conduct financial transactions, and allow entry into secure areas. They also provide directions for navigation, and a variety of other services.
- Most portable computers do not have keyboards.
- Rotating memories such as hard drives, CD-ROMs, and DVDs are on their way out.
- Most users have servers on their homes and offices where they keep large stores of digital objects, including, among other things, virtual reality environments, although these are still on an early stage.
- Cables are disappearing.
- The majority of text is created using continuous speech recognition, or CSR (dictation software). CSRs are very accurate, far more than the human transcriptionists, who were used up until a few years ago.

- Books, magazines, and newspapers are now routinely read on displays that are the size of small books.
- Computer displays built into eyeglasses are also used. These specialized glasses allow the users to see the normal environment while creating a virtual image that appears to hover in front of the viewer.
- Computers routinely include moving-picture image cameras and are able to reliably identify their owners from their faces.
- Three-dimensional chips are commonly used.
- Students from all ages have a portable computer, very thin and soft, weighting less than one pound. They interact with their computers primarily by voice and by pointing with a device that looks like a pencil. Keyboards still exist but most textual language is created by speaking.
- Intelligent courseware has emerged as a common means of learning; recent controversial studies have shown that students can learn basic skills such as reading and math just as readily with interactive learning software as with human teachers.
- Schools are increasingly relying on software approaches. Many children learn to read on their own using personal computers before entering grade school.
- Persons with disabilities are rapidly overcoming their handicaps through intelligent technology.

- Students with reading disabilities routinely use print-to-speech reading systems.
- Print-to-speech reading machines for the blind are now very small, inexpensive, palm-size devices that can read books.
- Useful navigation systems have finally been developed to assist blind people in moving and avoiding obstacles. Those systems use GPS technology. The blind person communicates with his navigation system by voice.
- Deaf persons commonly use portable speech-to-text listening machines which display a real-time transcription of what people are saying. The deaf user has the choice of either reading the transcribed speech as displayed text or watching an animated person gesturing in sign language.
- Listening machines can also translate what is being said into another language in real time, so they are commonly used by hearing people as well.
- There is a growing perception that the primary disabilities of blindness, deafness, and physical impairment do not necessarily [qualify as such]. Disabled persons routinely describe their disabilities as mere inconveniences.
- In communications, telephone translation technology is commonly used. This allow you to speak in English, while your Japanese friend hears you in Japanese, and vice versa.
- Telephones are primarily wireless and include high-resolution moving images.

- Haptic technologies are emerging. They allow people to touch and feel objects and other persons at a distance. These force-feedback devices are wildly used in games and in training simulation systems. Interactive games routinely include all-encompassing all-visual and auditory environments.
- The 1999 chat rooms have been replaced with virtual environments.
- At least half of all transactions are conducted online.
- Intelligent routes are in use, primarily for long-distance travel. Once your car's computer's guiding system locks on to the control sensors on one of these highways, you can sit back and relax.
- There is a growing neo-Luddite movement.

Now, just to be clear, I don't want you to look at all that and think, "Gee, the future goes more slowly than expected—technological progress must be naturally slow."

More like, "Where are you pulling all these burdensome details from, anyway?"

If you looked at all that and said, "Ha ha, how wrong; now I have my *own* amazing prediction for what the future will be like, *it won't be like that*," then you're really missing the whole "you have to work a whole lot harder to produce veridical beliefs about the future, and often the info you want is simply not obtainable" business.

* * *

Robin Hanson

It might be useful to put a little check or X mark next to these items, to indicate which were right vs. wrong, so the eye could quickly scan down the list to see the overall trend. But yes, it won't look good for Kurzweil, and checking such track records is very important.

Robin Hanson

In order to score forecasts, what we really want is:

1. Probabilities assigned to each item
2. Some other forecast of the same things to compare with

Without these we are stuck trying to guess what probability he had in mind and what probabilities others would have assigned back then to these same items.

See original post for all comments.

* * *

1. forever freedom, "My Disappointment at the Future," Longecity forum, July 26, 2007, accessed July 28, 2013, http : / / www . longecity . org / forum / topic / 17025 - my - disappointment-at-the-future/

   Quoted with minor changes to spelling and grammar.

31

I Heart Cyc

Robin Hanson

1 December 2008

Eliezer Tuesday:

> . . . Eurisko may *still* be the most sophisticated self-improving AI ever built—in the 1980s, by Douglas Lenat before he started wasting his life on Cyc. . . .
>
> Eurisko lacked what I called "insight"—that is, the type of abstract knowledge that lets humans fly through the search space.

I commented:

> [You] ignore that Lenat has his own theory which he gives as the *reason* he's been pursuing Cyc. You should at least explain why you think his theory wrong; I find his theory quite plausible.

Eliezer replied only:

> Artificial Addition, The Nature of Logic, Truly Part of You, Words as Mental Paintbrush Handles, Detached Lever Fallacy . . .

The main relevant points from these Eliezer posts seem to be that AI researchers wasted time on messy *ad hoc* nonmonotonic logics, while elegant mathy Bayes net approaches work much better; that it is much better to know how to generate specific knowledge from general principles than to just be told lots of specific knowledge; and that our minds have lots of hidden machinery behind the words we use; words as "detached levers" won't work. But I doubt Lenat or the Cyc folks disagree with any of these points.

The lesson Lenat took from EURISKO is that architecture is overrated; AIs learn slowly now mainly because they know so little. So we need to explicitly code knowledge by hand until we have enough to build systems effective at asking questions, reading, and learning for themselves. Prior AI researchers were too comfortable starting every project over from scratch; they needed to join to create larger integrated knowledge bases. This still seems to me a reasonable view, and anyone who thinks Lenat created the best AI system ever should consider seriously the lesson he thinks he learned.

Of course the Cyc project is open to criticism on its many particular choices. People have complained about its logic-like and language-like representations, about its selection of prototypical cases to build from (e.g., encyclopedia articles), about its focus on answering over acting, about how often it rebuilds vs. maintaining legacy systems, and about being private vs. publishing everything.

But any large project like this would produce such disputes, and it is not obvious any of its choices have been seriously wrong. They had to start somewhere, and in my opinion they have now collected a knowledge base with a truly spectacular size, scope, and integration.

Other architectures may well work better, but if knowing lots is anywhere near as important as Lenat thinks, I'd expect serious AI attempts to import Cyc's knowledge, translating it into a new representation. No other source has anywhere near Cyc's size, scope, and integration. But if so, how could Cyc be such a waste?

Architecture being overrated would make architecture-based fooms less plausible. Given how small a fraction of our commonsense knowledge it seems to have so far, Cyc gives little cause for optimism for human-level AI anytime soon. And as long as a system like Cyc is limited to taking no actions other than drawing conclusions and asking questions, it is hard to see it could be that dangerous, even if it knew a whole awful lot. (Influenced by an email conversation with Stephen Reed.)

Added: Guha and Lenat in '93:

> . . . The Cyc project . . . is *not* an experiment whose sole purpose is to test a hypothesis, . . . rather it is an engineering effort, aimed at constructing an artifact. . . . The artifact we are building is a shared information resource, which many programs can usefully draw upon. Ultimately, it may suffice to be *the* shared resource . . .
>
> If there is a central assumption behind Cyc, it has to do with Content being the bottleneck or chokepoint to achieving AI. I.e., you can get just so far twiddling with . . . empty AIR (Architecture, Implementation, Representation.) Sooner or later, someone has to bite the Content bullet. . . . The

> Implementation is just scaffolding to facilitate the accretion of that Content. . . . Our project has been driven continuously and exclusively by Content. I.e., we built and refined code only when we had to. I.e., as various assertions or behaviors weren't readily handled by the then-current implementation, those needs for additional representational expressiveness or efficiency led to changes or new features in the Cyc representation language or architecture.[1]

At the bottom of this page is a little box showing random OpenCyc statements "in its best English"; click on any concept to see more.[2] OpenCyc is a public subset of Cyc.

* * *

Eliezer Yudkowsky

So my genuine, actual reaction to seeing this post title was, "You heart *WHAT?*"

Knowledge isn't being able to repeat back English statements. This is true even of humans. It's a hundred times more true of AIs, even if you turn the words into tokens and put the tokens in tree structures.

A basic exercise to perform with any supposed AI is to replace all the English names with random gensyms and see what the AI can still do, if anything. Deep Blue remains invariant under this exercise. Cyc, maybe, could count—it may have a genuine understanding of the word "four"—and could check certain uncomplicatedly structured axiom sets for logical consistency, although not, of course, anything on the order of say Peano arithmetic. The rest of Cyc is bogus. If it knows about anything, it only knows about certain relatively small and simple mathematical objects, certainly nothing about the real world.

You can't get knowledge into a computer that way. At all. Cyc is composed almost entirely of fake knowledge (barring anything it knows about certain simply structured mathematical objects).

As a search engine or something, Cyc might be an interesting startup, though I certainly wouldn't invest in it. As an Artificial General Intelligence, Cyc is just plain awful. It's not just that most of it is composed of suggestively named LISP tokens, there are also the other hundred aspects of cognition that are simply entirely missing. Like, say, probabilistic reasoning, or decision theory, or sensing or acting or—

—for the love of Belldandy! How can you even call this sad little thing an AGI project?

So long as they maintained their current architecture, I would have no fear of Cyc even if there were a million programmers working on it and they had access to a computer the size of a moon, any more than I would live in fear of a dictionary program containing lots of words.

Cyc is so unreservedly hopeless, especially by comparison to EURISKO that came before it, that it makes me seriously wonder if Lenat is doing something that I'm not supposed to postulate because it can always be more simply explained by foolishness rather than conspiracy.

Of course there are even sillier projects. Hugo de Garis and Mentifex both come to mind.

Robin Hanson

. . . Conversation *is* action. Replacing every word you spoke or heard with a new random gensym would destroy your ability to converse with others. So that would be a terrible way to test your true knowledge that enables your conversation. I'll grant that an ability to converse is a limited ability, and the ability to otherwise act effectively greatly expands one's capability and knowledge.

Eliezer Yudkowsky

Okay . . . look at it this way. Chimpanzees share 95% of our DNA and have much of the same gross cytoarchitecture of their brains. You cannot explain

to *chimpanzees* that Paris is the capital of France. You can train them to hold up a series of signs saying "Paris," then "Is-Capital-Of," then "France." But you cannot explain to them that Paris is the capital of France.

And a chimpanzee's cognitive architecture is *hugely* more sophisticated than Cyc's. Cyc isn't close. It's not in the ballpark. It's not in the galaxy holding the star around which circles the planet whose continent contains the country in which lies the city that built the ballpark.

Robin Hanson

Eliezer, we can make computers do lots of things we can't train chimps to do. Surely we don't want to limit AI research to only achieving chimp behaviors. We want to be opportunistic—developing whatever weak abilities have the best chance of leading later to stronger abilities. Answering encyclopedia questions might be the best weak ability to pursue first. Or it might not. Surely we just don't know, right?

See original post for all comments.

* * *

1. R. V. Guha and Douglas B. Lenat, "Re: CycLing Paper Reviews," *Artificial Intelligence* 61, no. 1 (1993): 149–174, doi:10.1016/0004-3702(93)90100-P.

2. http://sw.opencyc.org/; dead page, redirects to OpenCyc project.

32

Is the City-ularity Near?

Robin Hanson

9 February 2010

The land around New York City is worth a *lot*. A 2008 analysis[1] estimated prices for land, not counting buildings etc., for four boroughs of the city plus nearby parts of New Jersey (2,770 square miles, equivalent to a fifty-two-mile square). The total land value for this area (total land times average price) was $5.5 trillion in 2002 and $28 trillion in 2006.

The Economist said that in 2002 all developed-nation real estate was worth $62 trillion.[2] Since raw land value is on average about a third[3] of total real-estate value, that puts New York-area real estate at over 30% of all developed-nation real estate in 2002! Whatever the exact number, clearly this agglomeration contains vast value.

New York land is valuable mainly because of how it is organized. People want to be there because they want to interact with other peo-

ple they expect to be there, and they expect those interactions to be quite mutually beneficial. If you could take any other fifty-mile square (of which Earth has seventy-two thousand) and create that same expectation of mutual value from interactions, you could get people to come there, make buildings, etc., and you could sell that land for many trillions of dollars of profit.

Yet the organization of New York was mostly set long ago based on old tech (e.g., horses, cars, typewriters). Worse, no one really understands at a deep level how it is organized or why it works so well. Different people understand different parts, in mostly crude empirical ways.

So what will happen when super-duper smarties wrinkle their brows so hard that out pops a deep mathematical theory of cities, explaining clearly how city value is produced? What if they apply their theory to designing a city structure that takes best advantage of our most advanced techs, of 7gen phones, twitter-pedias, flying Segways, solar panels, gene-mod pigeons, and super-fluffy cupcakes? Making each city aspect more efficient makes the city more attractive, increasing the gains from making other aspects more efficient, in a grand spiral of bigger and bigger gains.

Once they convince the world of the vast value in their super-stupendous city design, won't everyone flock there and pay mucho trillions for the privilege? Couldn't they leverage this lead into better theories, enabling better designs giving far more trillions, and then spend all that on a super-designed war machine based on those same super-insights, and turn us all into down dour super-slaves? So isn't the very mostest importantest cause ever to make sure that we, the friendly freedom fighters, find this super-deep city theory first?

Well, no, it isn't. We don't believe in a city-ularity because we don't believe in a super-city theory found in a big brain flash of insight. What makes cities work well is mostly getting lots of details right. Sure, new-tech-based city designs can work better, but gradual tech gains mean no city is suddenly vastly better than others. Each change has costs to be weighed against hoped-for gains. Sure, costs of change might be lower when making a whole new city from scratch, but for that to work you have to be damn sure you know which changes are actually good ideas.

For similar reasons, I'm skeptical of a blank-slate AI mind-design intelligence explosion. Sure, if there were a super mind theory that allowed vast mental efficiency gains all at once—but there isn't. Minds are vast complex structures full of parts that depend intricately on each other, much like the citizens of a city. Minds, like cities, best improve gradually, because you just never know enough to manage a vast redesign of something with such complex interdependent adaptations.

* * *

See original post for all comments.

* * *

1. Andrew Haughwout, James Orr, and David Bedoll, "The Price of Land in the New York Metropolitan Area," *Current Issues in Economics and Finance* 13, no. 3 (2008), accessed June 21, 2013, http://www.newyorkfed.org/research/current_issues/ci14-3/ci14-3.html.

2. "House of Cards," *The Economist*, May 29, 2003, http://www.economist.com/node/1794873.

3. Richard W. Douglas Jr., "Site Value Taxation and Manvel's Land Value Estimates," *American Journal of Economics and Sociology* 37, no. 2 (1978): 217–223, http://www.jstor.org/stable/3486442.

33

Recursive Self-Improvement

Eliezer Yudkowsky

1 December 2008

Followup to: Life's Story Continues, Surprised by Brains, Cascades, Cycles, Insight, . . . Recursion, Magic, Engelbart: Insufficiently Recursive, Total Nano Domination

I think that, at some point in the development of Artificial Intelligence, we are likely to see a *fast, local* increase in capability—"AI go FOOM." Just to be clear on the claim, "fast" means on a timescale of weeks or hours rather than years or decades; and "FOOM" means way the hell smarter than anything else around, capable of delivering in short time periods technological advancements that would take humans decades, probably including full-scale molecular nanotechnology (that it gets by, e.g., ordering custom proteins over the

Internet with seventy-two-hour turnaround time). Not, "ooh, it's a little Einstein but it doesn't have any robot hands, how cute."

Most people who object to this scenario object to the "fast" part. Robin Hanson objected to the "local" part. I'll try to handle both, though not all in one shot today.

We are setting forth to analyze the developmental velocity of an Artificial Intelligence. We'll break down this velocity into optimization slope, optimization resources, and optimization efficiency. We'll need to understand cascades, cycles, insight, and recursion; and we'll stratify our recursive levels into the metacognitive, cognitive, metaknowledge, knowledge, and object levels.

Quick review:

- "Optimization slope" is the goodness and number of opportunities in the volume of solution space you're currently exploring, on whatever your problem is.
- "Optimization resources" is how much computing power, sensory bandwidth, trials, etc. you have available to explore opportunities.
- "Optimization efficiency" is how well you use your resources. This will be determined by the goodness of your current mind design—the point in mind-design space that is your current self—along with its knowledge and metaknowledge (see below).

Optimizing *yourself* is a special case, but it's one we're about to spend a lot of time talking about.

By the time any mind solves some kind of *actual problem*, there's actually been a huge causal lattice of optimizations applied—for example, human brains evolved, and then humans developed the idea of science, and then applied the idea of science to generate knowledge about gravity, and then you use this knowledge of gravity to finally design a damn bridge or something.

So I shall stratify this causality into levels—the boundaries being semi-arbitrary, but you've got to draw them somewhere:

- "Metacognitive" is the optimization that builds the brain—in the case of a human, natural selection; in the case of an AI, either human programmers or, after some point, the AI itself.
- "Cognitive," in humans, is the labor performed by your neural circuitry, algorithms that consume large amounts of computing power but are mostly opaque to you. You know what you're seeing, but you don't know how the visual cortex works. The Root of All Failure in AI is to underestimate those algorithms because you can't see them . . . In an AI, the lines between procedural and declarative knowledge are theoretically blurred, but in practice it's often possible to distinguish cognitive algorithms and cognitive content.
- "Metaknowledge": Discoveries about how to discover, "Science" being an archetypal example, "Math" being another. You can think of these as reflective cognitive content (knowledge about how to think).
- "Knowledge": Knowing how gravity works.

- "Object level": Specific actual problems like building a bridge or something.

I am arguing that an AI's developmental velocity will not be smooth; the following are some classes of phenomena that might lead to non-smoothness. First, a couple of points that weren't raised earlier:

- *Roughness:* A search space can be naturally rough—have unevenly distributed *slope*. With constant optimization pressure, you could go through a long phase where improvements are easy, then hit a new volume of the search space where improvements are tough. Or vice versa. Call this factor *roughness*.
- *Resource overhangs:* Rather than resources growing incrementally by reinvestment, there's a big bucket o' resources behind a locked door, and once you unlock the door you can walk in and take them all.

And these other factors previously covered:

- *Cascades* are when one development leads the way to another—for example, once you discover gravity, you might find it easier to understand a coiled spring.
- *Cycles* are feedback loops where a process's output becomes its input on the next round. As the classic example of a fission chain reaction illustrates, a cycle whose underlying processes are continuous may show qualitative changes of surface behavior—a threshold of criticality—the difference between

each neutron leading to the emission of 0.9994 additional neutrons versus each neutron leading to the emission of 1.0006 additional neutrons. The effective neutron multiplication factor is k and I will use it metaphorically.

- *Insights* are items of knowledge that tremendously decrease the cost of solving a wide range of problems—for example, once you have the calculus insight, a whole range of physics problems become a whole lot easier to solve. Insights let you fly through, or teleport through, the solution space, rather than searching it by hand—that is, "insight" represents knowledge about the structure of the search space itself.

And finally:

- *Recursion* is the sort of thing that happens when you hand the AI the object-level problem of "redesign your own cognitive algorithms."

Suppose I go to an AI programmer and say, "Please write me a program that plays chess." The programmer will tackle this using their existing knowledge and insight in the domain of chess and search trees; they will apply any metaknowledge they have about how to solve programming problems or AI problems; they will process this knowledge using the deep algorithms of their neural circuitry; and this neutral circuitry will have been designed (or rather its wiring algorithm designed) by natural selection.

If you go to a sufficiently sophisticated AI—more sophisticated than any that currently exists—and say, "write me a chess-playing

program," the same thing might happen: The AI would use its knowledge, metaknowledge, and existing cognitive algorithms. Only the AI's *metacognitive* level would be, not natural selection, but the *object level* of the programmer who wrote the AI, using *their* knowledge and insight, etc.

Now suppose that instead you hand the AI the problem, "Write a better algorithm than X for storing, associating to, and retrieving memories." At first glance this may appear to be just another object-level problem that the AI solves using its current knowledge, metaknowledge, and cognitive algorithms. And indeed, in one sense it should be just another object-level problem. But it so happens that the AI itself uses algorithm X to store associative memories, so if the AI can improve on this algorithm, it can rewrite its code to use the new algorithm X+1.

This means that the AI's *metacognitive* level—the optimization process responsible for structuring the AI's cognitive algorithms in the first place—has now collapsed to identity with the AI's *object* level.

For some odd reason, I run into a lot of people who vigorously deny that this phenomenon is at all novel; they say, "Oh, humanity is already self-improving, humanity is already going through a FOOM, humanity is already in an Intelligence Explosion," etc., etc.

Now to me, it seems clear that—at this point in the game, in advance of the observation—it is *pragmatically* worth drawing a distinction between inventing agriculture and using that to support more professionalized inventors, versus directly rewriting your own source code in RAM. Before you can even *argue* about whether the two phenomena are likely to be similar in practice, you need to accept that they are, in fact, two different things to be argued *about*.

And I do expect them to be very distinct in practice. Inventing science is not rewriting your neural circuitry. There is a tendency to *completely overlook* the power of brain algorithms, because they are invisible to introspection. It took a long time historically for people to realize that there *was* such a thing as a cognitive algorithm that could underlie thinking. And then, once you point out that cognitive algorithms exist, there is a tendency to tremendously underestimate them, because you don't know the specific details of how your hippocampus is storing memories well or poorly—you don't know how it could be improved, or what difference a slight degradation could make. You can't draw detailed causal links between the wiring of your neural circuitry and your performance on real-world problems. All you can *see* is the knowledge and the metaknowledge, and that's where all your causal links go; that's all that's *visibly* important.

To see the brain circuitry vary, you've got to look at a chimpanzee, basically. Which is not something that most humans spend a lot of time doing, because chimpanzees can't play our games.

You can also see the tremendous overlooked power of the brain circuitry by observing what happens when people set out to program what looks like "knowledge" into Good-Old-Fashioned AIs, semantic nets and such. Roughly, nothing happens. Well, research papers happen. But no actual intelligence happens. Without those opaque, overlooked, invisible brain algorithms, there is no real knowledge—only a tape recorder playing back human words. If you have a small amount of fake knowledge, it doesn't do anything, and if you have a huge amount of fake knowledge programmed in at huge expense, it still doesn't do anything.

So the cognitive level—in humans, the level of neural circuitry and neural algorithms—is a level of tremendous but invisible power. The difficulty of penetrating this invisibility and creating a real cognitive level is what stops modern-day humans from creating AI. (Not that an AI's cognitive level would be made of neurons or anything equivalent to neurons; it would just do cognitive labor on the same level of organization.[1] Planes don't flap their wings, but they have to produce lift somehow.)

Recursion that can rewrite the cognitive level is *worth distinguishing*.

But to some, having a term so narrow as to refer to an AI rewriting its own source code, and not to humans inventing farming, seems hardly open, hardly embracing, hardly communal; for we all know that to say two things are similar shows greater enlightenment than saying that they are different. Or maybe it's as simple as identifying "recursive self-improvement" as a term with positive affective valence, so you figure out a way to apply that term to humanity, and then you get a nice dose of warm fuzzies. Anyway.

So what happens when you start rewriting cognitive algorithms?

Well, we do have *one* well-known historical case of an optimization process writing cognitive algorithms to do further optimization; this is the case of natural selection, our alien god.

Natural selection seems to have produced a pretty smooth trajectory of more sophisticated brains over the course of hundreds of millions of years. That gives us our first data point, with these characteristics:

- Natural selection on sexual multicellular eukaryotic life can probably be treated as, to first order, an optimizer of *roughly constant efficiency and constant resources*.
- Natural selection does not have anything akin to insights. It does sometimes stumble over adaptations that prove to be surprisingly reusable outside the context for which they were adapted, but it doesn't fly through the search space like a human. Natural selection is just *searching the immediate neighborhood of its present point in the solution space, over and over and over.*
- Natural selection *does* have cascades: adaptations open up the way for further adaptations.

So—*if* you're navigating the search space via the ridiculously stupid and inefficient method of looking at the neighbors of the current point, without insight—with constant optimization pressure—then . . .

Well, I've heard it claimed that the evolution of biological brains has accelerated over time, and I've also heard that claim challenged. If there's actually been an acceleration, I would tend to attribute that to the "adaptations open up the way for further adaptations" phenomenon—the more brain genes you have, the more chances for a mutation to produce a new brain gene. (Or, more complexly: The more organismal error-correcting mechanisms the brain has, the more likely a mutation is to produce something useful rather than fatal.) In the case of hominids in particular over the last few million years, we may also have been experiencing accelerated *selection* on

brain proteins, *per se*—which I would attribute to sexual selection, or brain variance accounting for a greater proportion of total fitness variance.

Anyway, what we definitely do *not* see under these conditions is *logarithmic* or *decelerating* progress. It did *not* take ten times as long to go from *H. erectus* to *H. sapiens* as from *H. habilis* to *H. erectus*. Hominid evolution did *not* take eight hundred million years of additional time, after evolution immediately produced *Australopithecus*-level brains in just a few million years after the invention of neurons themselves.

And another, similar observation: human intelligence does *not* require a hundred times as much computing power as chimpanzee intelligence. Human brains are merely three times too large, and our prefrontal cortices six times too large, for a primate with our body size.

Or again: It does not seem to require a thousand times as many genes to build a human brain as to build a chimpanzee brain, even though human brains can build toys that are a thousand times as neat.

Why is this important? Because it shows that with *constant optimization pressure* from natural selection and *no intelligent insight*, there were *no diminishing returns* to a search for better brain designs up to at least the human level. There were probably *accelerating* returns (with a low acceleration factor). There are no *visible speed bumps*, so far as I know.

But all this is to say only of natural selection, which is not recursive.

If you have an investment whose output is not coupled to its input—say, you have a bond, and the bond pays you a certain amount

of interest every year, and you spend the interest every year—then this will tend to return you a linear amount of money over time. After one year, you've received $10; after two years, $20; after three years, $30.

Now suppose you *change* the qualitative physics of the investment, by coupling the output pipe to the input pipe. Whenever you get an interest payment, you invest it in more bonds. Now your returns over time will follow the curve of compound interest, which is exponential. (Please note: *Not all accelerating processes are smoothly exponential.* But this one happens to be.)

The first process grows at a rate that is linear over *time*; the second process grows at a rate that is linear in its *cumulative return so far*.

The too-obvious mathematical idiom to describe the impact of recursion is replacing an equation

$$y = f(t)$$

with

$$\frac{dy}{dt} = f(y).$$

For example, in the case above, reinvesting our returns transformed the *linearly* growing

$$y = m \cdot t$$

into

$$\frac{dy}{dt} = m \cdot y$$

whose solution is the exponentially growing

$$y = e^{m \cdot t}.$$

Now . . . I do not think you can *really* solve equations like this to get anything like a description of a self-improving AI.

But it's the obvious reason why I *don't* expect the future to be a continuation of past trends. The future contains a feedback loop that the past does not.

As a different Eliezer Yudkowsky wrote, very long ago: "If computing power doubles every eighteen months, what happens when computers are doing the research?"[2]

And this sounds horrifyingly naive to my present ears, because that's not really how it works at all—but still, it illustrates the idea of "the future contains a feedback loop that the past does not."

History up until this point was a long story about natural selection producing humans, and then, after humans hit a certain threshold, humans starting to rapidly produce knowledge and metaknowledge that could—among other things—feed more humans and support more of them in lives of professional specialization.

To a first approximation, natural selection held still during human cultural development. Even if Gregory Clark's crazy ideas (Wikipedia) are crazy enough to be true—i.e., some human populations evolved lower discount rates and more industrious work habits over the course of just a few hundred years from 1200 to 1800[3]—that's just tweaking a few relatively small parameters; it is not the same as developing new complex adaptations with lots of interdependent parts. It's not a chimp-human type gap.

So then, *with human cognition remaining more or less constant*, we found that knowledge feeds off knowledge with $k > 1$—given a background of roughly constant cognitive algorithms at the human level. We discovered major chunks of metaknowledge, like Science

and the notion of Professional Specialization, that changed the exponents of our progress; having lots more humans around, due to, e.g., the object-level innovation of farming, may have have also played a role. Progress in any one area tended to be choppy, with large insights leaping forward, followed by a lot of slow incremental development.

With history *to date*, we've got a series of integrals looking something like this:

- Metacognitive = natural selection, optimization efficiency/resources roughly constant
- Cognitive = Human intelligence = integral of evolutionary optimization velocity over a few hundred million years, then roughly *constant* over the last ten thousand years
- Metaknowledge = Professional Specialization, Science, etc. = integral over cognition we did about procedures to follow in thinking, where metaknowledge can also feed on itself, there were major insights and cascades, etc.
- Knowledge = all that actual science, engineering, and general knowledge accumulation we did = integral of cognition + metaknowledge (current knowledge) over time, where knowledge feeds upon itself in what seems to be a roughly exponential process
- Object level = stuff we actually went out and did = integral of cognition + metaknowledge + knowledge (current solutions); over a short timescale this tends to be smoothly exponential to the degree that the people involved understand the idea of

> investments competing on the basis of interest rate, but over medium-range timescales the exponent varies, and on a long range the exponent seems to increase

If you were to summarize that in one breath, it would be, "With constant natural selection pushing on brains, progress was linear or mildly accelerating; with constant brains pushing on metaknowledge and knowledge and object-level progress feeding back to metaknowledge and optimization resources, progress was exponential or mildly superexponential."

Now fold back the object level so that it becomes the metacognitive level.

And note that we're doing this through a chain of differential equations, not just one; it's the *final* output at the object level, after all those integrals, that becomes the velocity of metacognition.

You should get . . .

. . . very fast progress? Well, no, not necessarily. You can also get nearly *zero* progress.

If you're a recursified optimizing compiler, you rewrite yourself just once, get a single boost in speed (like 50% or something), and then never improve yourself any further, ever again.

If you're EURISKO, you manage to modify some of your metaheuristics, and the metaheuristics work noticeably better, and they even manage to make a few further modifications to themselves, but then the whole process runs out of steam and flatlines.

It was human intelligence that produced these artifacts to begin with. Their *own* optimization power is far short of human—so incredibly weak that, after they push themselves along a little, they can't

push any further. Worse, their optimization at any given level is characterized by a limited number of opportunities, which once used up are gone—extremely sharp diminishing returns.

When you fold a complicated, choppy, cascade-y chain of differential equations in on itself via recursion, *it should either flatline or blow up*. You would need *exactly the right law of diminishing returns* to fly through the extremely narrow *soft-takeoff keyhole*.

The *observed history of optimization to date* makes this *even more unlikely*. I don't see any reasonable way that you can have constant evolution produce human intelligence on the observed historical trajectory (linear or accelerating), and constant human intelligence produce science and technology on the observed historical trajectory (exponential or superexponential), and *fold that in on itself*, and get out something whose rate of progress is in any sense *anthropomorphic*. From our perspective it should either flatline or FOOM.

When you first build an AI, it's a baby—if it had to improve *itself*, it would almost immediately flatline. So you push it along using your own cognition, metaknowledge, and knowledge—*not* getting any benefit of recursion in doing so, just the usual human idiom of knowledge feeding upon itself and insights cascading into insights. Eventually the AI becomes sophisticated enough to start improving *itself*, not just small improvements, but improvements large enough to cascade into other improvements. (Though right now, due to lack of human insight, what happens when modern researchers push on their AGI design is mainly nothing.) And then you get what I. J. Good called an "intelligence explosion."

I even want to say that the functions and curves being such as to allow hitting the soft-takeoff keyhole is *ruled out* by observed history

to date. But there are small conceivable loopholes, like "maybe all the curves change drastically and completely as soon as we get past the part we know about in order to give us exactly the right anthropomorphic final outcome," or "maybe the trajectory for insightful optimization of intelligence has a law of diminishing returns where blind evolution gets accelerating returns."

There's other factors contributing to hard takeoff, like the existence of hardware overhang in the form of the poorly defended Internet and fast serial computers. There's more than one possible species of AI we could see, given this whole analysis. I haven't yet touched on the issue of localization (though the basic issue is obvious: the initial recursive cascade of an intelligence explosion can't race through human brains because human brains are not modifiable until the AI is already superintelligent).

But today's post is already too long, so I'd best continue tomorrow.

Post scriptum: It occurred to me just after writing this that I'd been victim of a cached Kurzweil thought in speaking of the knowledge level as "exponential." Object-level resources are exponential in human history because of physical cycles of reinvestment. If you try defining knowledge as productivity per worker, I expect that's exponential too (or productivity growth would be unnoticeable by now as a component in economic progress). I wouldn't be surprised to find that published journal articles are growing exponentially. But I'm not quite sure that it makes sense to say humanity has learned as much since 1938 as in all earlier human history . . . though I'm quite willing to believe we produced more goods . . . then again we surely learned more since 1500 than in all the time before. Anyway, human knowledge being "exponential" is a more complicated issue than I made it

out to be. But the human object level is more clearly exponential or superexponential.

* * *

Robin Hanson

Depending on which abstractions you emphasize, you can describe a new thing as something completely new under the sun, or as yet another example of something familiar. So the issue is which abstractions make the most sense to use. We have seen cases before where when one growth via some growth channel opened up more growth channels to further enable growth. So the question is how similar those situations are to this situation, where an AI getting smarter allows an AI to change its architecture in more and better ways. Which is another way of asking which abstractions are most relevant.

Eliezer Yudkowsky

. . . Well, the whole post above is just putting specific details on that old claim, "Natural selection producing humans and humans producing technology can't be extrapolated to an AI insightfully modifying its low-level brain algorithms, because the latter case contains a feedback loop of an importantly different type; it's like trying to extrapolate a bird flying outside the atmosphere or extrapolating the temperature/compression law of a gas past the point where the gas becomes a black hole."

If you just pick an abstraction that isn't detailed enough to talk about the putative feedback loop, and then insist on extrapolating out the old trends from the absence of the feedback loop, I would consider this a weak response. . . .

See original post for all comments.

* * *

1. Eliezer Yudkowsky, "Levels of Organization in General Intelligence," in *Artificial General Intelligence*, ed. Ben Goertzel and Cassio Pennachin, Cognitive Technologies (Berlin: Springer, 2007), doi:10.1007/978-3-540-68677-4, 389–501.

2. Eliezer Yudkowsky, "Staring into the Singularity" (Unpublished manuscript, 1996), last revised May 27, 2001, http://yudkowsky.net/obsolete/singularity.html.

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

34

Whither Manufacturing?

Robin Hanson

2 December 2008

Back in the ’70s many folks thought they knew what the future of computing looked like: everyone sharing time slices of a few huge computers. After all, they saw that CPU cycles, the main computing cost, were cheaper on bigger machines. This analysis, however, ignored large administrative overheads in dealing with shared machines. People eagerly grabbed personal computers (PCs) to avoid those overheads, even though PC CPU cycles were more expensive.

Similarly, people seem to make lots of assumptions when they refer to “full-scale nanotechnology.” This phrase seems to elicit images of fridge-sized home appliances that, when plugged in and stocked with a few “toner cartridges,” make anything a CAD system can describe, and so quickly and cheaply that only the most price-sensitive folks would consider making stuff any other way. It seems people

learned too much from the PC case, thinking everything must become personal and local. (Note computing is now getting *less* local.) But *there is no general law of increasingly local production*.

The locality of manufacturing, and computing as well, have always come from tradeoffs between economies and diseconomies of scale. Things can often be made cheaper in big centralized plants, especially if located near key inputs. When processing bulk materials, for example, there is a rough two-thirds-cost power law: throughput goes as volume, while the cost to make and manage machinery tends to go as surface area. But it costs more to transport products from a few big plants. Local plants can offer more varied products, explore more varied methods, and deliver cheaper and faster.

Innovation and adaption to changing conditions can be faster or slower at centralized plants, depending on other details. Politics sometimes pushes for local production to avoid dependence on foreigners, and at other times pushes for central production to make succession more difficult. Smaller plants can better avoid regulation, while larger ones can gain more government subsidies. When formal intellectual property is weak (the usual case), producers can prefer to make and sell parts instead of selling recipes for making parts.

Often producers don't even really know how they achieve the quality they do. Manufacturers today make great use of expensive intelligent labor; while they might prefer to automate all production, they just don't know how. It is not at all obvious how feasible is "full nanotech," if defined as fully automated manufacturing, in the absence of full AI. Nor is it obvious that even fully automated manufacturing would be very local production. The optimal locality will depend on how all these factors change over the coming decades; don't

be fooled by confident conclusions based on only one or two of these factors. More here.[1]

* * *

Eliezer Yudkowsky

I have no objection to most of this—the main thing that I think deserves pointing out is the idea that you can serve quite a lot of needs by having "nanoblocks" that reconfigure themselves in response to demands. I'd think this would be a localizing force with respect to production, and a globalizing force with respect to design.

Robin Hanson

Eliezer, the less local is manufacturing, the harder it will be for your super-AI to build undetected the physical equipment it needs to take over the world.

Eliezer Yudkowsky

Robin, a halfway transhuman social intelligence should have *no trouble* coming up with good excuses or bribes to cover nearly anything it wants to do. We're not talking about grey goo here, we're talking about something that can invent its own cover stories. Current protein synthesis machines are not local—most labs send out to get the work done, though who knows how long that will stay true—but I don't think it would be very difficult for a smart AI to use them "undetected," that is, without any alarms sounding about the order placed.

Robin Hanson

Eliezer, it might take more than a few mail-order proteins to take over the world. . . .

Eliezer Yudkowsky

. . . Robin, why does it realistically take more than a few mail-order proteins to take over the world? Ribosomes are reasonably general molecular factories and quite capable of self-replication to boot.

Robin Hanson

Eliezer, I guess I'm just highlighting the extreme degree of intelligence postulated, that this week-old box that has made no visible outside mark beyond mail-ordering a few proteins knows enough to use those proteins to build a physically small manufacturing industry that is more powerful than the entire rest of the world.

Eliezer Yudkowsky

Ergh, just realized that I didn't do a post discussing the bogosity of "human-equivalent computing power" calculations. Well, here's a start in a quick comment—Moravec, in 1988, used Moore's Law to calculate how much power we'd have in 2008.[2] He more or less nailed it. He spent a lot of pages justifying the idea that Moore's Law could continue, but from our perspective that seems more or less prosaic.

Moravec spent fewer pages than he did on Moore's Law justifying his calculation that the supercomputers we would have in 2008 would be "human-equivalent brainpower."

Did Moravec nail that as well? Given the sad state of AI theory, we actually have no evidence against it. But personally, I suspect that he overshot; I suspect that one could build a mind of formidability roughly comparable to

human on a modern-day desktop computer, or maybe even a desktop computer from 1996; because I now think that evolution wasn't all that clever with our brain design, and that the 100 Hz serial speed limit on our neurons has to be having all sorts of atrocious effects on algorithmic efficiency. If it was a superintelligence doing the design, you could probably have roughly human formidability on something substantially smaller.

Just a very rough eyeball estimate, no real numbers behind it.

See original post for all comments.

* * *

1. Hanson, "Five Nanotech Social Scenarios."

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

35

Hard Takeoff

Eliezer Yudkowsky

2 December 2008

Continuation of: Recursive Self-Improvement

Constant natural selection pressure, operating on the genes of the hominid line, produced improvement in brains over time that seems to have been, roughly, *linear or accelerating*; the operation of constant human brains on a pool of knowledge seems to have produced returns that are, very roughly, *exponential or superexponential*. (Robin proposes that human progress is well characterized as a series of exponential modes with diminishing doubling times.[1])

Recursive self-improvement (RSI)—an AI rewriting its own cognitive algorithms—identifies the object level of the AI with a force acting on the metacognitive level; it "closes the loop" or "folds the graph in on itself." E.g., the difference between returns on a constant

investment in a bond and reinvesting the returns into purchasing further bonds is the difference between the equations $y = f(t) = m \cdot t$ and $\frac{dy}{dt} = f(y) = m \cdot y$, whose solution is the compound interest exponential $y = e^{m \cdot t}$.

When you fold a whole chain of differential equations in on itself like this, it should either peter out rapidly as improvements fail to yield further improvements, or else go FOOM. An *exactly right law of diminishing returns* that lets the system fly through the *soft-takeoff keyhole* is unlikely—*far* more unlikely than seeing such behavior in a system with a roughly constant underlying optimizer, like evolution improving brains, or human brains improving technology. Our present life is no good indicator of things to come.

Or to try and compress it down to a slogan that fits on a T-shirt—not that I'm saying this is a good idea—"Moore's Law is exponential *now*; it would be really odd if it *stayed* exponential with the improving computers *doing the research*." I'm not saying you literally get $\frac{dy}{dt} = e^y$ that goes to infinity after finite time—and hardware improvement is in some ways the least interesting factor here—but should we really see the same curve we do now?

RSI is the biggest, most interesting, hardest-to-analyze, sharpest break with the past contributing to the notion of a "hard takeoff" a.k.a. "AI go FOOM," but it's nowhere near being the *only* such factor. The advent of human intelligence was a discontinuity with the past even *without* RSI . . .

. . . which is to say that observed evolutionary history—the discontinuity between humans and chimps, who share 95% of our DNA—*lightly* suggests a critical threshold built into the capabilities that we

think of as "general intelligence," a machine that becomes far more powerful once the last gear is added.

This is only a *light* suggestion because the branching time between humans and chimps *is* enough time for a good deal of complex adaptation to occur. We could be looking at the sum of a cascade, not the addition of a final missing gear. On the other hand, we can look at the gross brain anatomies and see that human brain anatomy and chimp anatomy have not diverged all that much. On the gripping hand, there's the sudden cultural revolution—the sudden increase in the sophistication of artifacts—that accompanied the appearance of anatomically modern Cro-Magnons just a few tens of thousands of years ago.

Now of course this might all just be completely inapplicable to the development trajectory of AIs built by human programmers rather than by evolution. But it at least *lightly suggests*, and provides a hypothetical *illustration* of, a discontinuous leap upward in capability that results from a natural feature of the solution space—a point where you go from sorta-okay solutions to totally amazing solutions as the result of a few final tweaks to the mind design.

I could potentially go on about this notion for a bit—because, in an evolutionary trajectory, it can't *literally* be a "missing gear," the sort of discontinuity that follows from removing a gear that an otherwise functioning machine was built around. So if you suppose that a final set of changes was enough to produce a sudden huge leap in effective intelligence, it does demand the question of what those changes were. Something to do with reflection—the brain modeling or controlling itself—would be one obvious candidate. Or perhaps a change in motivations (more curious individuals, using the brainpower they have

in different directions) in which case you *wouldn't* expect that discontinuity to appear in the AI's development, but you would expect it to be more effective at earlier stages than humanity's evolutionary history would suggest . . . But you could have whole journal issues about that one question, so I'm just going to leave it at that.

Or consider the notion of sudden resource bonanzas. Suppose there's a semi-sophisticated Artificial General Intelligence running on a cluster of a thousand CPUs. The AI has not hit a wall—it's still improving itself—but its self-improvement is going so *slowly* that, the AI calculates, it will take another fifty years for it to engineer/implement/refine just the changes it currently has in mind. Even if this AI would go FOOM eventually, its current progress is so slow as to constitute being flatlined . . .

So the AI turns its attention to examining certain blobs of binary code—code composing operating systems, or routers, or DNS services—and then takes over all the poorly defended computers on the Internet. This may not require what humans would regard as genius, just the ability to examine lots of machine code and do relatively low-grade reasoning on millions of bytes of it. (I have a saying/hypothesis that a *human* trying to write *code* is like someone without a visual cortex trying to paint a picture—we can do it eventually, but we have to go pixel by pixel because we lack a sensory modality for that medium; it's not our native environment.) The Future may also have more legal ways to obtain large amounts of computing power quickly.

This sort of resource bonanza is intriguing in a number of ways. By assumption, optimization *efficiency* is the same, at least for the moment—we're just plugging a few orders of magnitude more resource into the current input/output curve. With a stupid algorithm,

a few orders of magnitude more computing power will buy you only a linear increase in performance—I would not fear Cyc even if it ran on a computer the size of the Moon, because there is no there there.

On the other hand, humans have a brain three times as large, and a prefrontal cortex six times as large, as that of a standard primate our size—so with software improvements of the sort that natural selection made over the last five million years, it does not require exponential increases in computing power to support linearly greater intelligence. Mind you, this sort of biological analogy is always fraught—maybe a human has not much more cognitive horsepower than a chimpanzee, the same underlying tasks being performed, but in a few more domains and with greater reflectivity—the engine outputs the same horsepower, but a few gears were reconfigured to turn each other less wastefully—and so you wouldn't be able to go from human to superhuman with just another sixfold increase in processing power . . . or something like that.

But if the lesson of biology suggests anything, it is that you do not run into logarithmic returns on *processing power* in the course of reaching human intelligence, even when that processing power increase is strictly parallel rather than serial, provided that you are at least as good as writing software to take advantage of that increased computing power as natural selection is at producing adaptations—five million years for a sixfold increase in computing power.

Michael Vassar observed in yesterday's comments that humans, by spending linearly more time studying chess, seem to get linear increases in their chess rank (across a wide range of rankings), while putting exponentially more time into a search algorithm is usually required to yield the same range of increase. Vassar called this "bizarre,"

but I find it quite natural. Deep Blue searched the raw game tree of chess; Kasparov searched the compressed regularities of chess. It's not surprising that the simple algorithm gives logarithmic returns and the sophisticated algorithm is linear. One might say similarly of the course of human progress seeming to be closer to exponential, while evolutionary progress is closer to being linear. Being able to understand the regularity of the search space counts for quite a lot.

If the AI is somewhere in between—not as brute-force as Deep Blue, nor as compressed as a human—then maybe a ten-thousandfold increase in computing power will only buy it a tenfold increase in optimization velocity . . . but that's still quite a speedup.

Furthermore, all *future* improvements the AI makes to itself will now be amortized over ten thousand times as much computing power to apply the algorithms. So a single improvement to *code* now has more impact than before; it's liable to produce more further improvements. Think of a uranium pile. It's always running the same "algorithm" with respect to neutrons causing fissions that produce further neutrons, but just piling on more uranium can cause it to go from subcritical to supercritical, as any given neutron has more uranium to travel through and a higher chance of causing future fissions.

So just the resource bonanza represented by "eating the Internet" or "discovering an application for which there is effectively unlimited demand, which lets you rent huge amounts of computing power while using only half of it to pay the bills"—even though this event isn't particularly *recursive* of itself, just an object-level fruit-taking—could potentially drive the AI from subcritical to supercritical.

Not, mind you, that this will happen with an AI that's just stupid. But an AI already improving itself *slowly*—that's a different case.

Even if this doesn't happen—if the AI uses this newfound computing power at all effectively, its optimization efficiency will increase more quickly than before—just because the AI has *more* optimization power to apply to the task of increasing its own efficiency, thanks to the sudden bonanza of optimization resources.

So the *whole trajectory* can conceivably change, just from so simple and straightforward and unclever and uninteresting-seeming an act as eating the Internet. (Or renting a bigger cloud.)

Agriculture changed the course of human history by supporting a larger population—and that was just a question of having more humans around, not individual humans having a brain a hundred times as large. This gets us into the whole issue of the returns on scaling individual brains not being anything like the returns on scaling the number of brains. A big-brained human has around four times the cranial volume of a chimpanzee, but four chimps $\neq$ one human. (And for that matter, sixty squirrels $\neq$ one chimp.) Software improvements here almost certainly completely dominate hardware, of course. But having a thousand scientists who collectively read all the papers in a field, and who talk to each other, is not like having one superscientist who has read all those papers and can correlate their contents directly using native cognitive processes of association, recognition, and abstraction. Having more humans talking to each other using low-bandwidth words cannot be expected to achieve returns similar to those from scaling component cognitive processes within a coherent cognitive system.

This, too, is an idiom outside human experience—we *have* to solve big problems using lots of humans, because there is no way to solve them using ONE BIG human. But it never occurs to anyone to substi-

tute four chimps for one human; and only a certain very foolish kind of boss thinks you can substitute ten programmers with one year of experience for one programmer with ten years of experience.

(Part of the general Culture of Chaos that praises emergence, and thinks evolution is smarter than human designers, also has a mythology of groups being inherently superior to individuals. But this is generally a matter of poor individual rationality, and various arcane group structures that are supposed to compensate, rather than an inherent fact about cognitive processes somehow *scaling better when chopped up into distinct brains*. If that were *literally* more efficient, evolution would have designed humans to have four chimpanzee heads that argued with each other. In the realm of AI, it seems much more straightforward to have a single cognitive process that lacks the emotional stubbornness to cling to its accustomed theories, and doesn't *need* to be argued out of it at gunpoint or replaced by a new generation of grad students. I'm not going to delve into this in detail for now, just warn you to be suspicious of this particular creed of the Culture of Chaos; it's not like they actually *observed* the relative performance of a hundred humans versus one BIG mind with a brain fifty times human size.)

So yes, there was a lot of software improvement involved—what we are seeing with the modern human brain size, is probably not so much the brain volume *required* to support the software improvement, but rather the *new evolutionary equilibrium* for brain size *given* the improved software.

Even so—hominid brain size increased by a factor of five over the course of around five million years. You might want to think *very seriously* about the contrast between that idiom, and a successful AI being

able to expand onto five thousand times as much hardware over the course of five minutes—when you are pondering possible hard takeoffs, and whether the AI trajectory ought to look similar to human experience.

A subtler sort of hardware overhang, I suspect, is represented by modern CPUs having a 2 GHz *serial speed*, in contrast to neurons that spike a hundred times per second on a good day. The "hundred-step rule" in computational neuroscience is a rule of thumb that any postulated neural algorithm which runs in real time has to perform its job in less than one hundred *serial* steps one after the other.[2] We do not understand how to efficiently use the computer hardware we have now to do intelligent thinking. But the much-vaunted "massive parallelism" of the human brain is, I suspect, mostly cache lookups to make up for the sheer awkwardness of the brain's *serial* slowness—if your computer ran at 200 Hz, you'd have to resort to all sorts of absurdly massive parallelism to get anything done in real time. I suspect that, if *correctly designed*, a midsize computer cluster would be able to get high-grade thinking done at a serial speed much faster than human, even if the total parallel computing power was less.

So that's another kind of overhang: because our computing hardware has run so far ahead of AI *theory*, we have incredibly fast computers we don't know how to use *for thinking*; getting AI *right* could produce a huge, discontinuous jolt, as the speed of high-grade thought on this planet suddenly dropped into computer time.

A still subtler kind of overhang would be represented by human failure to use our gathered experimental data efficiently.

On to the topic of insight, another potential source of discontinuity: The course of hominid evolution was driven by evolution's neigh-

borhood search; if the evolution of the brain accelerated to some degree, this was probably due to existing adaptations creating a greater number of possibilities for further adaptations. (But it couldn't accelerate past a certain point, because evolution is limited in how much selection pressure it can apply—if someone succeeds in breeding due to adaptation A, that's less variance left over for whether or not they succeed in breeding due to adaptation B.)

But all this is searching the raw space of genes. Human design intelligence, or sufficiently sophisticated AI design intelligence, isn't like that. One might even be tempted to make up a completely different curve out of thin air—like, intelligence will take all the easy wins first, and then be left with only higher-hanging fruit, while increasing complexity will defeat the ability of the designer to make changes. So where blind evolution accelerated, intelligent design will run into diminishing returns and grind to a halt. And as long as you're making up fairy tales, you might as well further add that the law of diminishing returns will be exactly right, and have bumps and rough patches in exactly the right places, to produce a smooth gentle takeoff even after recursion and various hardware transitions are factored in . . . One also wonders why the story about "intelligence taking easy wins first in designing brains" *tops out* at or before human-level brains, rather than going *a long way beyond human* before topping out. But one suspects that if you tell *that* story, there's no point in inventing a law of diminishing returns to begin with.

(Ultimately, if the character of physical law is anything like our current laws of physics, there will be limits to what you can do on finite hardware, and limits to how much hardware you can assemble

in finite time, but if they are very *high* limits relative to human brains, it doesn't affect the basic prediction of hard takeoff, "AI go FOOM.")

The main thing I'll venture into actually expecting from adding "insight" to the mix, is that there'll be a discontinuity at the point where the AI *understands how to do AI theory*, the same way that human researchers try to do AI theory. An AI, to swallow its own optimization chain, must not just be able to rewrite its own source code; it must be able to, say, rewrite *Artificial Intelligence: A Modern Approach* (2nd Edition). An ability like this seems (untrustworthily, but I don't know what else to trust) like it ought to appear at around the same time that the architecture is at the level of, or approaching the level of, being able to handle what humans handle—being no shallower than an actual human, whatever its inexperience in various domains. It would produce further discontinuity at around that time.

In other words, when the AI becomes smart enough to *do AI theory*, that's when I expect it to fully swallow its own optimization chain and for the *real* FOOM to occur—though the AI might *reach* this point as part of a cascade that started at a more primitive level.

All these complications are why I don't believe we can *really* do any sort of math that will predict *quantitatively* the trajectory of a hard takeoff. You can make up models, but real life is going to include all sorts of discrete jumps, bottlenecks, bonanzas, insights—and the "fold the curve in on itself" paradigm of recursion is going to amplify even small roughnesses in the trajectory.

So I stick to qualitative predictions. "AI go FOOM."

Tomorrow I hope to tackle locality, and a bestiary of some possible qualitative trajectories the AI might take given this analysis. Robin Hanson's summary of "primitive AI fooms to sophisticated AI"

doesn't fully represent my views—that's just one entry in the bestiary, albeit a major one.

* * *

See original post for all comments.

* * *

1. Robin Hanson, "Long-Term Growth as a Sequence of Exponential Modes" (Unpublished manuscript, 1998), last revised December 2000, http : / / hanson . gmu . edu / longgrow.pdf.

2. J. A. Feldman and Dana H. Ballard, "Connectionist Models and Their Properties," *Cognitive Science* 6, no. 3 (1982): 205–254, doi:10.1207/s15516709cog0603_1.

36

Test Near, Apply Far

Robin Hanson

3 December 2008

Companies often ask me if prediction markets can forecast distant future topics. I tell them yes, but that is not the place to test any doubts about prediction markets. To vet or validate prediction markets, you want topics where there will be many similar forecasts over a short time, with other mechanisms making forecasts that can be compared.

If you came up with an account of the cognitive processes that allowed Newton or Einstein to make their great leaps of insight, you would want to look for where that, or related accounts, applied to more common insight situations. An account that only applied to a few extreme "geniuses" would be much harder to explore, since we know so little about those few extreme cases.

If you wanted to explain the vast voids we seem to see in the distant universe, and you came up with a theory of a new kind of mat-

ter that could fill that void, you would want to ask where nearby one might find or be able to create that new kind of matter. Only after confronting this matter theory with local data would you have much confidence in applying it to distant voids.

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.

* * *

Eliezer Yudkowsky

Considering the historical case of the advent of human intelligence, how would you have wanted to handle it using only abstractions that could have been tested before human intelligence showed up?

(This being one way of testing your abstraction about abstractions . . .)

We recently had a cute little "black swan" in our financial markets. It wasn't really very black. But some people predicted it well enough to make money off it, and some people didn't. Do you think that someone could have triumphed

using your advice here, with regards to that particular event which is now near to us? If so, how?

Robin Hanson

Eliezer, it is very hard to say what sort of other experience and evidence there would have been "near" hypothetical creatures who know of Earth history before humans, to guess if that evidence would have been enough to guide them to good abstractions to help them anticipate and describe the arrival of humans. For some possible creatures, they may well not have had enough to do a decent job.

See original post for all comments.

37

Permitted Possibilities and Locality

Eliezer Yudkowsky

3 December 2008

Continuation of: Hard Takeoff

The analysis given in the last two days permits more than one possible AI trajectory:

1. Programmers, smarter than evolution at finding tricks that work, but operating without fundamental insight or with only partial insight, create a mind that is dumber than the researchers but performs lower-quality operations much faster. This mind reaches $k > 1$, cascades up to the level of a very smart human, *itself* achieves insight into intelligence, and undergoes the really fast part of the FOOM, to superintelligence.

This would be the major nightmare scenario for the origin of an unFriendly AI.

2. Programmers operating with partial insight create a mind that performs a number of tasks very well, but can't really handle self-modification let alone AI theory. A mind like this might progress with something like smoothness, pushed along by the researchers rather than itself, even all the way up to average-human capability—not having the insight into its own workings to push itself any further. We also suppose that the mind is either already using huge amounts of available hardware, or scales *very* poorly, so it cannot go FOOM just as a result of adding a hundred times as much hardware. This scenario seems less likely to my eyes, but it is not *ruled out* by any effect I can see.

3. Programmers operating with strong insight into intelligence directly create, along an efficient and planned pathway, a mind capable of modifying itself with deterministic precision—provably correct or provably noncatastrophic self-modifications. This is the only way I can see to achieve narrow enough targeting to create a Friendly AI. The "natural" trajectory of such an agent would be slowed by the requirements of precision, and sped up by the presence of insight; but because this is a Friendly AI, notions like "You can't yet improve yourself this far, your goal system isn't verified enough" would play a role.

So these are some things that I think are permitted to happen, albeit that case (2) would count as a hit against me to some degree because it does seem unlikely.

Here are some things that *shouldn't* happen, on my analysis:

- An *ad hoc* self-modifying AI as in (1) undergoes a cycle of self-improvement, starting from stupidity, that carries it up to the level of a very smart human—and then stops, unable to progress any further. (The upward slope in this region is supposed to be very steep!)
- A mostly non-self-modifying AI as in (2) is pushed by its programmers up to a roughly human level . . . then to the level of a very smart human . . . then to the level of a mild transhuman . . . but the mind still does not achieve insight into its own workings and still does not undergo an intelligence explosion—just continues to increase smoothly in intelligence from there.

And I also don't think this is allowed: the "scenario that Robin Hanson seems to think is the line-of-maximum-probability for AI as heard and summarized by Eliezer Yudkowsky":

- No one AI that does everything humans do, but rather a large, diverse population of AIs. These AIs have various *domain-specific* competencies that are "human+ level"—not just in the sense of Deep Blue beating Kasparov, but in the sense that, in these domains, the AIs seem to have good "common sense" and can, e.g., recognize, comprehend and handle situations that weren't in their original programming. But only in the special domains for which that AI was crafted/trained. Collectively,

these AIs may be strictly more competent than any one human, but no individual AI is more competent than any one human.

- Knowledge and even skills are widely traded in this economy of AI systems.
- In concert, these AIs, and their human owners, and the economy that surrounds them, undergo a *collective* FOOM of self-improvement. No local agent is capable of doing all this work, only the collective system.
- The FOOM's benefits are distributed through a whole global economy of trade partners and suppliers, including existing humans and corporations, though existing humans and corporations may form an increasingly small fraction of the New Economy.
- This FOOM looks like an exponential curve of compound interest, like the modern world but with a substantially shorter doubling time.

Mostly, Robin seems to think that uploads will come first, but that's a whole 'nother story. So far as AI goes, this looks like Robin's maximum line of probability—and if I got this mostly wrong or all wrong, that's no surprise. Robin Hanson did the same to me when summarizing what he thought were my own positions. I have never thought, in prosecuting this Disagreement, that we were starting out with a mostly good understanding of what the Other was thinking; and this seems like an important thing to have always in mind.

So—bearing in mind that I may well be criticizing a straw misrepresentation, and that I know this full well, but I am just trying to guess

my best—here's what I see as wrong with the elements of this scenario:

The abilities we call "human" are the final products of an economy of mind—not in the sense that there are selfish agents in it, but in the sense that there are production lines; and I would even expect evolution to enforce something approaching fitness as a common unit of currency. (Enough selection pressure to create an adaptation from scratch should be enough to fine-tune the resource curves involved.) It's the production lines, though, that are the main point—that your brain has specialized parts and the specialized parts pass information around. All of this goes on behind the scenes, but it's what finally *adds up* to any *single* human ability.

In other words, trying to get humanlike performance in *just one* domain is divorcing a final product of that economy from all the work that stands behind it. It's like having a global economy that can *only* manufacture toasters, but not dishwashers or light bulbs. You can have something like Deep Blue that beats humans at chess in an inhuman, specialized way; but I don't think it would be easy to get humanish performance at, say, biology R&D, without a whole mind and architecture standing behind it that would also be able to accomplish other things. Tasks that draw on our cross-domain-ness, or our long-range real-world strategizing, or our ability to formulate new hypotheses, or our ability to use very high-level abstractions—I don't think that you would be able to replace a human in just that one job, without also having something that would be able to learn many different jobs.

I think it is a fair analogy to the idea that you shouldn't see a global economy that can manufacture toasters but not manufacture anything else.

This is why I don't think we'll see a system of AIs that are diverse, individually highly specialized, and *only collectively* able to do anything a human can do.

Trading cognitive content around between diverse AIs is more difficult and less likely than it might sound. Consider the field of AI as it works today. Is there *any* standard database of cognitive content that you buy off the shelf and plug into your amazing new system, whether it be a chess player or a new data-mining algorithm? If it's a chess-playing program, there are databases of stored games—but that's not the same as having databases of preprocessed cognitive content.

So far as I can tell, the diversity of cognitive architectures acts as a *tremendous* barrier to trading around cognitive content. If you have many AIs around that are all built on the same architecture by the same programmers, they might, *with a fair amount of work*, be able to pass around learned cognitive content. Even this is less trivial than it sounds. If two AIs both see an apple for the first time, and they both independently form concepts about that apple, and they both independently build some new cognitive content around those concepts, then their *thoughts* are effectively written in a different language. By seeing a single apple at the same time, they could identify a concept they both have in mind, and in this way build up a common language . . .

. . . the point being that, even when two separated minds are running literally the same source code, it is still difficult for them to trade new knowledge *as raw cognitive content* without having a special language designed just for sharing knowledge.

Now suppose the two AIs are built around different architectures.

The barrier this opposes to a true, cross-agent, literal "economy of mind" is so strong that, in the vast majority of AI applications you set out to write today, you will not bother to import any standardized preprocessed cognitive content. It will be easier for your AI application to start with some standard examples—databases of *that* sort of thing do exist, in some fields anyway—and *redo all the cognitive work of learning* on its own.

That's how things stand today.

And I have to say that, looking over the diversity of architectures proposed at any AGI conference I've attended, it is very hard to imagine directly trading cognitive content between any two of them. It would be an immense amount of work just to set up a language in which they could communicate what they take to be facts about the world—never mind preprocessed cognitive content.

This is a force for *localization*: unless the condition I have just described changes drastically, it means that agents will be able to do their own cognitive labor, rather than needing to get their brain content manufactured elsewhere, or even being *able* to get their brain content manufactured elsewhere. I can imagine there being an exception to this for *non*-diverse agents that are deliberately designed to carry out this kind of trading within their code-clade. (And in the long run, difficulties of translation seems less likely to stop superintelligences.)

But in *today's* world, it seems to be the rule that when you write a new AI program, you can sometimes get preprocessed raw data, but you will not buy any preprocessed cognitive content—the internal content of your program will come from within your program.

And it actually does seem to me that AI would have to get *very* sophisticated before it got over the "hump" of increased sophistication making sharing harder instead of easier. I'm not sure this is pre-takeoff sophistication we're talking about, here. And the cheaper computing power is, the easier it is to just share the *data* and do the *learning* on your own.

Again—in today's world, sharing of cognitive content between diverse AIs doesn't happen, even though there are lots of machine learning algorithms out there doing various jobs. You could say things would happen differently in the future, but it'd be up to you to make that case.

Understanding the difficulty of interfacing diverse AIs is the next step toward understanding why it's likely to be a *single coherent* cognitive system that goes FOOM via recursive self-improvement. The same sort of barriers that apply to trading direct cognitive content would also apply to trading changes in cognitive source code.

It's a whole lot easier to modify the source code in the interior of your own mind than to take that modification and sell it to a friend who happens to be written on different source code.

Certain kinds of abstract insights would be more tradeable, among sufficiently sophisticated minds; and the major insights might be well worth selling—like, if you invented a new *general* algorithm at

some subtask that many minds perform. But if you again look at the modern state of the field, then you find that it is only a few algorithms that get any sort of general uptake.

And if you hypothesize minds that understand these algorithms, and the improvements to them, and what these algorithms are for, and how to implement and engineer them—then these are already very sophisticated minds; at this point, they are AIs that can do their own AI theory. So the hard takeoff has to have not already started, yet, at this point where there are many AIs around that can do AI theory. If they can't do AI theory, diverse AIs are likely to experience great difficulties trading code improvements among themselves.

This is another localizing force. It means that the improvements you make to yourself, and the compound interest earned on those improvements, are likely to stay local.

If the scenario with an AI takeoff is anything at all like the modern world in which all the attempted AGI projects have completely incommensurable architectures, then any self-improvements will definitely stay put, not spread.

But suppose that the situation *did* change drastically from today, and that you had a community of diverse AIs which were sophisticated enough to share cognitive content, code changes, and even insights. And suppose even that this is true at the *start* of the FOOM—that is, the community of diverse AIs got all the way up to that level, without yet using a FOOM or starting a FOOM at a time when it would still be localized.

We can even suppose that most of the code improvements, algorithmic insights, and cognitive content driving any particular AI are

coming from outside that AI—sold or shared—so that the improvements the AI makes to *itself* do not dominate its total velocity.

Fine. The *humans* are not out of the woods.

Even if we're talking about uploads, it will be immensely more difficult to apply any of the algorithmic insights that are tradeable between AIs to the undocumented human brain that is a huge mass of spaghetti code, that was never designed to be upgraded, that is not end-user-modifiable, that is not hot-swappable, that is written for a completely different architecture than what runs efficiently on modern processors . . .

And biological humans? Their neurons just go on doing whatever neurons do, at one hundred cycles per second (tops).

So this FOOM that follows from recursive self-improvement, the cascade effect of using your increased intelligence to rewrite your code and make yourself even smarter—

The barriers to sharing cognitive improvements among diversely designed AIs are large; the barriers to sharing with uploaded humans are incredibly huge; the barrier to sharing with biological humans is essentially absolute. (Barring a [benevolent] superintelligence with nanotechnology, but if one of those is around, you have already won.)

In this hypothetical global economy of mind, the humans are like a country that no one can invest in, that cannot adopt any of the new technologies coming down the line.

I once observed that Ricardo's Law of Comparative Advantage is the theorem that unemployment should not exist. The gotcha being that if someone is sufficiently unreliable, there is a cost to you to train them, a cost to stand over their shoulders and monitor them, a cost to check their results for accuracy—the existence of unemployment in

our world is a combination of transaction costs like taxes, regulatory barriers like minimum wage, and above all, *lack of trust*. There are a dozen things I would pay someone else to do for me—if I wasn't paying taxes on the transaction, and if I could trust a stranger as much as I trust myself (both in terms of their honesty and of acceptable quality of output). Heck, I'd as soon have some formerly unemployed person walk in and spoon food into my mouth while I kept on typing at the computer—if there were no transaction costs, and I trusted them.

If high-quality thought drops into a speed closer to computer time by a few orders of magnitude, no one is going to take a subjective year to explain to a biological human an idea that they will be barely able to grasp, in exchange for an even slower guess at an answer that is probably going to be wrong anyway.

Even *uploads* could easily end up doomed by this effect, not just because of the immense overhead cost and slowdown of running their minds, but because of the continuing error-proneness of the human architecture. Who's going to trust a giant messy undocumented neural network, any more than you'd run right out and hire some unemployed guy off the street to come into your house and do your cooking?

This FOOM leaves humans behind . . .

. . . unless you go the route of Friendly AI, and make a superintelligence that simply *wants* to help humans, not for any economic value that humans provide to it, but because that is its nature.

And just to be clear on something—which really should be clear by now, from all my other writing, but maybe you're just wandering in—it's not that having squishy things running around on two legs is the ultimate height of existence. But if you roll up a random AI with

a random utility function, it just ends up turning the universe into patterns we would not find very eudaimonic—turning the galaxies into paperclips. If you try a haphazard attempt at making a "nice" AI, the sort of not-even-half-baked theories I see people coming up with on the spot and occasionally writing whole books about, like using reinforcement learning on pictures of smiling humans to train the AI to value happiness (yes, this was a book) then the AI just transforms the galaxy into tiny molecular smileyfaces . . .

It's not some small, mean desire to survive for myself, at the price of greater possible futures, that motivates me. The thing is—those greater possible futures, they don't happen automatically. There are stakes on the table that are so much an invisible background of your existence that it would never occur to you they could be lost; and these things will be shattered by default, if not specifically preserved.

And as for the idea that the whole thing would happen slowly enough for humans to have plenty of time to react to things—a smooth exponential shifted into a shorter doubling time—of that, I spoke yesterday. Progress seems to be exponential now, more or less, or at least accelerating, and that's with constant human brains. If you take a nonrecursive accelerating function and fold it in on itself, you are going to get superexponential progress. "If computing power doubles every eighteen months, what happens when computers are doing the research" should not just be a faster doubling time. (Though, that said, on any sufficiently short timescale progress might well *locally* approximate an exponential because investments will shift in such fashion that the marginal returns on investment balance, even in the interior of a single mind; interest rates consis-

tent over a timespan imply smooth exponential growth over that timespan.)

You can't count on warning, or time to react. If an accident sends a sphere of plutonium, not critical, but *prompt critical*, neutron output can double in a tenth of a second even with $k = 1.0006$. It can deliver a killing dose of radiation or blow the top off a nuclear reactor before you have time to draw a breath. Computers, like neutrons, already run on a timescale much faster than human thinking. We are already past the world where we can definitely count on having time to react.

When you move into the transhuman realm, you also move into the realm of adult problems. To wield great power carries a price in great precision. You can build a nuclear reactor but you can't ad-lib it. On the problems of this scale, if you want the universe to end up a worthwhile place, you can't just throw things into the air and trust to luck and later correction. That might work in childhood, but not on adult problems where the price of one mistake can be instant death.

Making it into the future is an adult problem. That's not a death sentence. I think. It's not the *inevitable* end of the world. I hope. But if you want human*kind* to survive, and the future to be a worthwhile place, then this will take careful crafting of the first superintelligence—not just letting economics or *whatever* take its easy, natural course. The easy, natural course is fatal—not just to ourselves but to all our hopes.

That, itself, is natural. It is only to be expected. To hit a narrow target you must aim; to reach a good destination you must steer; to win, you must make an extra-ordinary effort.

* * *

See original post for all comments.

38

Underconstrained Abstractions

Eliezer Yudkowsky

4 December 2008

Followup to: The Weak Inside View

Saith Robin:

> 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.

Well . . . I understand why one would have that reaction. But I'm not sure we can *really* get away with that.

When possible, I try to talk in concepts that can be verified with respect to existing history. When I talk about natural selection not running into a law of diminishing returns on genetic complexity or brain size, I'm talking about something that we can try to verify by looking at the capabilities of other organisms with brains big and small. When I talk about the boundaries to sharing cognitive content between AI programs, you can look at the field of AI the way it works today and see that, lo and behold, there isn't a lot of cognitive content shared.

But in my book this is just *one* trick in a *library* of methodologies for dealing with the Future, which is, in general, a hard thing to predict.

Let's say that instead of using my complicated-sounding disjunction (many *different* reasons why the growth trajectory might contain an upward cliff, which don't *all* have to be true), I instead staked my *whole* story on the critical threshold of human intelligence. Saying, "Look how sharp the slope is here!"—well, it would *sound* like a simpler story. It would be closer to fitting on a T-shirt. And by talking about *just* that one abstraction and no others, I could make it sound like I was dealing in verified historical facts—humanity's evolutionary history is something that has already happened.

But speaking of an abstraction being "verified" by previous history is a tricky thing. There is this little problem of *underconstraint*—of there being more than one possible abstraction that the data "verifies."

In "Cascades, Cycles, Insight" I said that economics does not seem to me to deal much in the origins of novel knowledge and novel designs, and said, "If I underestimate your power and merely parody

your field, by all means inform me what kind of economic study has been done of such things." This challenge was answered by comments directing me to some papers on "endogenous growth," which happens to be the name of theories that don't take productivity improvements as exogenous forces.

I've looked at some literature on endogenous growth. And don't get me wrong, it's probably not too bad as economics. However, the seminal literature talks about ideas being generated by combining other ideas, so that if you've got N ideas already and you're combining them three at a time, that's a potential $N!/((3!)(N-3)!)$ new ideas to explore. And then goes on to note that, in this case, there will be vastly more ideas than anyone can explore, so that the rate at which ideas are exploited will depend more on a paucity of explorers than a paucity of ideas.

Well . . . first of all, the notion that "ideas are generated by combining other ideas N at a time" is not exactly an amazing AI theory; it is an economist looking at, essentially, the whole problem of AI, and trying to solve it in five seconds or less. It's not as if any experiment was performed to actually watch ideas recombining. Try to build an AI around this theory and you will find out in very short order how useless it is as an account of where ideas come from . . .

But more importantly, if the only proposition you actually *use* in your theory is that there are more ideas than people to exploit them, then this is the only proposition that can even be *partially* verified by testing your theory.

Even if a recombinant growth theory can be fit to the data, then the historical data still underconstrains the *many* possible abstractions that might describe the number of possible ideas available—any

hypothesis that has around "more ideas than people to exploit them" will fit the same data equally well. You should simply say, "I assume there are more ideas than people to exploit them," not go so far into mathematical detail as to talk about N choose 3 ideas. It's not that the dangling math here is underconstrained by the *previous* data, but that you're not even using it *going forward*.

(And does it even fit the data? I have friends in venture capital who would laugh like hell at the notion that there's an unlimited number of really good ideas out there. Some kind of Gaussian or power-law or something distribution for the goodness of available ideas seems more in order . . . I don't object to "endogenous growth" simplifying things for the sake of having one simplified abstraction and seeing if it fits the data well; we all have to do that. Claiming that the underlying math doesn't *just* let you build a useful model, but *also* has a fairly direct correspondence to reality, ought to be a whole 'nother story, in economics—or so it seems to me.)

(If I merely misinterpret the endogenous growth literature or underestimate its sophistication, by all means correct me.)

The further away you get from highly regular things like atoms, and the closer you get to surface phenomena that are the final products of many moving parts, the more history underconstrains the abstractions that you use. This is part of what makes futurism difficult. If there were obviously only one story that fit the data, who would bother to use anything else?

Is Moore's Law a story about the increase in computing power *over time*—the number of transistors on a chip as a function of how far the planets have spun in their orbits, or how many times a light wave emitted from a cesium atom has changed phase?

Or does the same data equally verify a hypothesis about exponential increases in investment in manufacturing facilities and R&D, with an even higher exponent, showing a law of diminishing returns?

Or is Moore's Law showing the increase in computing power as a function of some kind of optimization pressure applied by human researchers, themselves thinking at a certain rate?

That last one might seem hard to verify, since we've never watched what happens when a chimpanzee tries to work in a chip R&D lab. But on some raw, elemental level—would the history of the world *really* be just the same, proceeding on *just exactly* the same timeline as the planets move in their orbits, if, for these last fifty years, the researchers themselves had been running on the latest generation of computer chip at any given point? That sounds to me even sillier than having a financial model in which there's no way to ask what happens if real estate prices go down.

And then, when you apply the abstraction going forward, there's the question of whether there's more than one way to apply it—which is one reason why a lot of futurists tend to dwell in great gory detail on the past events that seem to support their abstractions, but just *assume* a single application forward.

E.g., Moravec in '88, spending a lot of time talking about how much "computing power" the human brain seems to use—but much less time talking about whether an AI would use the same amount of computing power, or whether using Moore's Law to extrapolate the first supercomputer of this size is the right way to time the arrival of AI. (Moravec thought we were supposed to have AI around *now*, based on his calculations—and he *under*estimated the size of the supercomputers we'd actually have in 2008.[1])

That's another part of what makes futurism difficult—after you've told your story about the past, even if it seems like an abstraction that can be "verified" with respect to the past (but what if you overlooked an alternative story for the same evidence?) that often leaves a lot of slack with regards to exactly what will happen with respect to that abstraction, going forward.

So if it's not as simple as *just* using the one trick of finding abstractions you can easily verify on available data . . .

. . . what are some other tricks to use?

* * *

Robin Hanson

So what exactly are you concluding from the fact that a seminal model has some unrealistic aspects, and that the connection between models and data in this field is not direct? That this field is useless as a source of abstractions? That it is no more useful than any other source of abstractions? That your abstractions are just as good?

Robin Hanson

Eliezer, is there some existing literature that has found "natural selection not running into a law of diminishing returns on genetic complexity or brain size," or are these new results of yours? These would seem to me quite publishable, though journals would probably want to see a bit more analysis than you have shown us.

Eliezer Yudkowsky

Robin, for some odd reason, it seems that a lot of fields in a lot of areas just analyze the abstractions they need for their own business, rather than the ones that you would need to analyze a self-improving AI.

I don't know if anyone has previously asked whether natural selection runs into a law of diminishing returns. But I observe that the human brain is only four times as large as a chimp brain, not a thousand times as large. And that most of the architecture seems to be the same; but I'm not deep enough into that field to know whether someone has tried to determine whether there are a lot more genes involved. I do know that brain-related genes were under stronger positive selection in the hominid line, but not so much stronger as to imply that, e.g., a thousand times as much selection pressure went into producing human brains from chimp brains as went into producing chimp brains in the first place. This is good enough to carry my point.

I'm not picking on endogenous growth, just using it as an example. I wouldn't be at all surprised to find that it's a fine theory. It's just that, so far as I can tell, there's some math tacked on that isn't actually used for anything, but provides a causal "good story" that doesn't actually sound all that good if you happen to study idea generation on a more direct basis. I'm just using it to make the point—it's not enough for an abstraction to fit the data, to be "verified." One should actually be aware of how the data is *constraining* the abstraction. The recombinant growth notion is an example of an abstraction that fits, but isn't constrained. And this is a general problem in futurism.

If you're going to start criticizing the strength of abstractions, you should criticize your own abstractions as well. How constrained are they by the data, really? Is there more than one reasonable abstraction that fits the same data?

Talking about what a field uses as "standard" doesn't seem like a satisfying response. Leaving aside that this is also the plea of those whose financial models don't permit real estate prices to go down—"it's industry standard, everyone is doing it"—what's standard in one field may not be standard in another, and you should be careful when turning an old standard to a new purpose. Sticking with standard endogenous growth models would be one matter if you wanted to just look at a human economy investing a usual fraction of money in R&D, and another matter entirely if your real interest and major concern was how

ideas scale *in principle*, for the sake of doing new calculations on what happens when you can buy research more cheaply.

There's no free lunch in futurism—no simple rule you can follow to make sure that your own preferred abstractions will automatically come out on top.

Robin Hanson

Eliezer, the factor of four between human and chimp brains seems to be far from sufficient to show that natural selection doesn't hit diminishing returns. In general I'm complaining that you mainly seem to ask us to believe your own new unvetted theories and abstractions, while I try when possible to rely on abstractions developed in fields of research (e.g., growth theory and research policy) where hundreds of researchers have worked full-time for decades to make and vet abstractions, confronting them with each other and data. You say your new approaches are needed because this topic area is far from previous ones, and I say test near, apply far; there is no free lunch in vetting; unvetted abstractions cannot be trusted just because it would be convenient to trust them. Also, note you keep talking about "verify," a very high standard, whereas I talked about the lower standards of "vet and validate."

Eliezer Yudkowsky

Robin, suppose that 1970 was the year when it became possible to run a human-equivalent researcher in real time using the computers of that year. Would the further progress of Moore's Law have been different from that in our own world, relative to sidereal time? Which abstractions are you using to answer this question? Have they been vetted and validated by hundreds of researchers?

Robin Hanson

Eliezer, my "Economic Growth Given Machine Intelligence"[2] *does* use one of the simplest endogenous growth models to explore how Moore's Law changes with computer-based workers. It is an early and crude attempt, but it is the sort of approach I think promising.

Eliezer Yudkowsky

Robin, I just read through that paper. Unless I missed something, you do not discuss, or even mention as a possibility, the effect of having around minds that are *faster* than human. You're just making a supply of em labor *cheaper* over time due to Moore's Law *treated as an exogenous growth factor*. Do you see why I might not think that this model was *even remotely on the right track*?

So . . . to what degree would you call the abstractions in your model "standard" and "vetted"?

How many new assumptions, exactly, are fatal? How many new terms are you allowed to introduce into an old equation before it becomes "unvetted," a "new abstraction"?

And if I devised a model that was no *more* different from the standard—departed by no *more* additional assumptions—than this one, which described the effect of faster researchers, would it be just as good, in your eyes?

Because there's a very simple and obvious model of what happens when your researchers obey Moore's Law, which makes even fewer new assumptions, and adds fewer terms to the equations . . .

You understand that if we're to have a standard that excludes some new ideas as being too easy to make up, then—even if we grant this standard—it's very important to ensure that standard is being applied *evenhandedly*, and not just *selectively* to exclude models that arrive at the wrong conclusions, because only in the latter case does it seem "obvious" that the new model is "unvetted." Do you *know* the criterion—can you say it aloud for all to hear—that you use to determine whether a model is based on vetted abstractions?

Robin Hanson

. . . Eliezer, the simplest standard model of endogenous growth is "learning by doing," where productivity increases with quantity of practice. That is the approach I tried in my paper. Also, while economists have many abstractions for modeling details of labor teams and labor markets, our standard is that the simplest versions should be of just a single aggregate quantity of labor. This one parameter of course implicitly combines the number of workers, the number of hours each works, how fast each thinks, how well trained they are, etc. If you instead have a one-parameter model that only considers how fast each worker thinks, you must be implicitly assuming all these other contributions stay constant. When you have only a single parameter for a sector in a model, it is best if that single parameter is an aggregate intended to describe that entire sector, rather than a parameter of one aspect of that sector.

Eliezer Yudkowsky

If one woman can have a baby in nine months, nine women can have a baby in one month? Having a hundred times as many people does not seem to scale even close to the same way as the effect of working for a hundred times as many years. This is a thoroughly vetted truth in the field of software management.

In science, time scales as the cycle of picking the best ideas in each generation and building on them; population would probably scale more like the right end of the curve generating what will be the best ideas of that generation.

Suppose Moore's Law to be endogenous in research. If I have new research-running CPUs with a hundred times the speed, I can use that to run the same number of researchers a hundred times as fast, or I can use it to run a hundred times as many researchers, or any mix thereof which I choose. I will choose the mix that maximizes my speed, of course. So the effect has to be at *least* as strong as speeding up time by a factor of a hundred. If you want to use a labor model that gives results stronger than that, go ahead . . .

Robin Hanson

Eliezer, it would be reasonable to have a model where the research sector of labor had a different function for how aggregate quantity of labor varied with the speed of the workers. . . .

See original post for all comments.

* * *

1. Moravec, *Mind Children*.

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

39

Beware Hockey-Stick Plans

Robin Hanson

4 December 2008

Eliezer yesterday:

> So really, the whole hard takeoff analysis of "flatline or FOOM" just ends up saying, "the AI will not hit the human timescale keyhole." From our perspective, an AI will either be so slow as to be bottlenecked, or so fast as to be FOOM. When you look at it that way, it's not so radical a prediction, is it?

Dot-com business plans used to have infamous "hockey-stick" market projections, a slow start that soon "fooms" into the stratosphere. From "How to Make Your Business Plan the Perfect Pitch":

> Keep your market-size projections conservative and defend whatever numbers you provide. If you're in the very early stages, most likely you can't calculate an accurate market

> size anyway. Just admit that. Tossing out ridiculous hockey-stick estimates will only undermine the credibility your plan has generated up to this point.[1]

Imagine a business trying to justify its hockey-stick forecast:

> We analyzed a great many models of product demand, considering a wide range of possible structures and parameter values (assuming demand never shrinks, and never gets larger than world product). We found that almost all these models fell into two classes: slow cases where demand grew much slower than the interest rate, and fast cases where it grew much faster than the interest rate. In the slow class we basically lose most of our million-dollar investment, but in the fast class we soon have profits of billions. So in expected value terms, our venture is a great investment, even if there is only a 0.1% chance the true model falls in this fast class.

What is wrong with this argument? It is that we have seen very few million-dollar investments ever give billions in profits. Nations and species can also have very complex dynamics, especially when embedded in economies and ecosystems, but few ever grow a thousand-fold, or have long stretches of accelerating growth. And the vast silent universe also suggests explosive growth is rare. So we are rightly skeptical about hockey-stick forecasts, even if they in some sense occupy half of an abstract model space.

Eliezer seems impressed that he can think of many ways in which AI growth could be "recursive," i.e., where all else equal one kind of growth makes it easier, rather than harder, to grow in other ways. But standard growth theory has many situations like this. For example, rising populations have more people to develop innovations of

all sorts; lower transportation costs allow more scale economies over larger integrated regions for many industries; tougher equipment allows more kinds of places to be farmed, mined and colonized; and lower info storage costs allow more kinds of business processes to be studied, tracked, and rewarded. And note that new ventures rarely lack for coherent stories to justify their hockey-stick forecasts.

The strongest data suggesting that accelerating growth is possible for more than a short while is the overall accelerating growth seen in human history. But since that acceleration has actually been quite discontinuous, concentrated in three sudden growth-rate jumps, I'd look more for sudden jumps than continuous acceleration in future growth as well. And unless new info sharing barriers are closer to the human-chimp barrier than to the farming and industry barriers, I'd also expect worldwide rather than local jumps. (More to come on locality.)

* * *

Eliezer Yudkowsky

The vast majority of AIs *won't* hockey-stick. In fact, creating a good AI design appears to be even harder than creating Microsoft's business plan.

But it would seem that, in fact, some companies do successfully create really high demand for their products. That is, the hockey-stick projection comes true in some cases. So it can't be the case that there's a universal law of diminishing returns that would prevent Microsoft or Google from existing—no matter how many dot-com companies made stupid claims. Reversed stupidity is not intelligence.

If everyone wants to *claim* they'll get the hockey-stick, that's not too surprising. Lots of people want to claim they've got the True AI Design, too, but

that doesn't make the problem of intelligence any more intrinsically difficult; it is what it is.

Human economies have many kinds of diminishing returns stemming from poor incentives, organizational scaling, regulatory interference, increased taxation when things seem to be going well enough to get away with it, etc., which would not plausibly carry over to a single mind. What argument is there for *fundamentally* diminishing returns?

And the basic extrapolation from Moore's Law to "Moore's Law when computers are doing the research" just doesn't seem like something you could acceptably rely on. *Recursion* is not the same as *cascades*. This is not just that one thing leads to another. What was once a protected level exerting a constant pressure will putatively have the output pipe connected straight into it. The very nature of the curve should change, like the jump from owning one bond that makes regular payments to reinvesting the payments.

Robin Hanson

I'm not saying nothing ever explodes; I'm saying the mere ability to find models wherein an explosion happens says little about if it will actually happen.

Eliezer, grabbing low-hanging fruit first is a very fundamental cause of diminishing returns. You don't seem to accept my description of "recursion" as "where all else equal one kind of growth makes it easier, rather than harder, to grow in other ways." Can you offer a precise but differing definition? . . .

Eliezer Yudkowsky

A "recursive" version of a scenario differs from a "nonrecursive" one in that there is a new feedback loop, connecting the final output of a chain of one or more optimizations to the design and structural state of an optimization process close to the start of the chain.

E.g., instead of evolution making minds, there are minds making minds.

Robin Hanson

Eliezer, but in my "recursion" examples there are new feedback loops. For example, before transportation tech starts changing, the scale of interaction is limited, but after it starts changing interaction scales increase, allowing a more specialized economy, including more specialized transportation, which allows transportation tech to better evolve.

See original post for all comments.

* * *

1. Michael V. Copeland, "How to Make Your Business Plan the Perfect Pitch," *Business 2.0*, September 1, 2005, http://money.cnn.com/magazines/business2/business2_archive/2005/09/01/8356496/.

40

Evolved Desires

Robin Hanson

5 December 2008

To a first approximation, the future will either be a *singleton*, a single integrated power choosing the future of everything, or it will be *competitive*, with conflicting powers each choosing how to perpetuate themselves. Selection effects apply robustly to competition scenarios; some perpetuation strategies will tend to dominate the future. To help us choose between a singleton and competition, and between competitive variations, we can analyze selection effects to understand competitive scenarios. In particular, selection effects can tell us the key feature without which it is very hard to forecast: *what creatures want*.

This seems to me a promising place for mathy folks to contribute to our understanding of the future. Current formal modeling tech-

niques are actually up to this task, and theorists have already learned lots about evolved preferences:

Discount Rates: Sexually reproducing creatures discount reproduction-useful resources given to their half-relations (e.g., kids, siblings) at a rate of one-half relative to themselves. Since in a generation they get too old to reproduce, and then only half-relations are available to help, they discount time at a rate of one-half per generation. Asexual creatures do not discount this way, though both types discount in addition for overall population growth rates. This suggests a substantial advantage for asexual creatures when discounting is important.

Local Risk: Creatures should care about their lineage success, i.e., the total number of their gene's descendants, weighted perhaps by their quality and relatedness, but shouldn't otherwise care *which* creatures sharing their genes now produce those descendants. So they are quite tolerant of risks that are uncorrelated, or negatively correlated, within their lineage. But they can care a lot more about risks that are correlated across such siblings. So they can be terrified of global catastrophe, mildly concerned about car accidents, and completely indifferent to within-lineage tournaments.

Global Risk: The total number of descendants within a lineage, and the resources it controls to promote future reproduction, vary across time. How risk averse should creatures be about short-term fluctuations in these such totals? If long-term future success is directly linear in current success, so that having twice as much now gives twice as much in the distant future, all else equal, you might

think creatures would be completely risk-neutral about their success now. Not so. Turns out selection effects *robustly* prefer creatures who have logarithmic preferences over success now. On global risks, they are quite risk averse.

Carl Shulman disagrees, claiming risk-neutrality:

> For such entities utility will be close to linear with the fraction of the accessible resources in our region that are dedicated to their lineages. A lineage . . . destroying all other life in the Solar System before colonization probes could escape . . . would gain nearly the maximum physically realistic utility. . . . A 1% chance of such victory would be 1% as desirable, but equal in desirability to an even, transaction-cost free division of the accessible resources with 99 other lineages.[1]

When I pointed Carl to the literature,[2] he replied:

> The main proof about maximizing log growth factor in individual periods . . . involves noting that, if a lineage takes gambles involving a particular finite risk of extinction in exchange for an increased growth factor in that generation, the probability of extinction will go to 1 over infinitely many trials. . . . But I have been discussing a finite case, and with a finite maximum of possible reproductive success attainable within our Hubble Bubble, expected value will generally not climb to astronomical heights as the probability of extinction approaches 1. So I stand by the claim that a utility function with utility linear in reproductive success over a world history will tend to win out from evolutionary competition.[3]

Imagine creatures that cared only about their lineage's fraction of the Hubble volume in a trillion years. If total success over this time is the product of success factors for many short time intervals, then induced preferences over each factor quickly approach log as the number of factors gets large. This happens for a wide range of risk attitudes toward final success, as long as the factors are not perfectly correlated. (Technically, if $U(\prod_t^N r_t) = \sum_t^N u(r_t)$, most $U(x)$ give $u(x)$ near $\log(x)$ for N large.)

A battle for the solar system is only one of many events where a lineage could go extinct in the next trillion years; why should evolved creatures treat it differently? Even if you somehow knew that it was in fact that last extinction possibility forevermore, how could evolutionary selection have favored a different attitude toward such that event? There cannot have been a history of previous last extinction events to select against creatures with preferences poorly adapted to such events. Selection prefers log preferences over a wide range of timescales up to some point where selection gets quiet. For an intelligence (artificial or otherwise) inferring very long term preferences by abstracting from its shorter time preferences, the obvious option is log preferences over *all* possible timescales.

Added: To explain my formula $U(\prod_t^N r_t) = \sum_t^N u(r_t)$,

- $U(x)$ is your final preferences over resources/copies of x at the "end."
- r_t is the ratio by which your resources/copies increase in each time step.
- $u(r_t)$ is your preferences over the next time step.

The right-hand side is expressed in a linear form so that if probabilities and choices are independent across time steps, then to maximize U, you'd just pick r_t to max the expected value of $u(r_t)$. For a wide range of $U(x)$, $u(x)$ goes to $\log(x)$ for N large.

* * *

Carl Shulman

> If total success over this time is the product of success factors for many short time intervals . . . [a] battle for the solar system is only one of many events where a lineage could go extinct in the next trillion years; why should evolved creatures treat it differently?

What sort of factors are you thinking about for a singleton expanding into our limited and apparently uninhabited accessible region, with current physical limits (thermodynamics, no FTL, etc.) assumed? Are you thinking about the entities' credence in the hypothesis that resources can increase vastly beyond those that physical limits seem to suggest? If resources could grow indefinitely, e.g., if there was a technological way to circumvent the laws of thermodynamics, then entities with unbounded utility functions (whether linear or log in reproductive success) will all have their calculations dominated by that possibility, and avoid struggles in the solar system that reduce their chances of getting access to such unbounded growth. I'm planning to talk more about that, but I started off with an assumption of common knowledge of current physics to illustrate dynamics.

> There cannot have been a history of previous last extinction events to select against creatures with preferences poorly adapted to such events.

Intelligent, foresightful entities with direct preferences for total reproductive success will mimic whatever local preferences would do best in a particular situation, so they won't be selected against; but in any case where the environment

changes so that evolved local preferences are no longer optimal, those with direct preferences for total success will be able to adapt immediately, without mutation and selection.

Robin Hanson

Carl, you lost me. Your first quote of me isn't talking about a singleton, and I don't see how physics knowledge is relevant. On your response to your second quote of me, you can't just assume you know what sort of risk aversion regarding the final outcome is the "true" preferences for "total success." If evolution selects for log preferences on all timescales on which it acts, why isn't log risk aversion the "true" total-success risk aversion? . . .

Carl Shulman

I'll reply in a post.

Eliezer Yudkowsky

> Robin: If evolution selects for log preferences on all timescales on which it acts, why isn't log risk aversion the "true" total success risk aversion?

Entities with logarithmic preferences over their aggregate number of copies in total world-histories should behave sublogarithmically when making local, independent choices on the next generation. The evolutionary analysis similarly talks about entities that you are likely to see in the sense of their being most frequent, not entities whose logarithms you are likely to see.

You can't literally have logarithmic preferences at both global and local timescales, I think. If global preference is logarithmic, wouldn't local preference be log-log?

Anyway, would you agree that: a linear aggregate utility over *complete world-histories* corresponds to logarithmic choices over *spatially global, temporally local options*, whose outcome you believe to be *uncorrelated* to the outcome of similar choices in future times.

Robin Hanson

Eliezer, I think you are just mistaken; log preferences aggregate or split in time to log preferences. Regarding your last question, I said a wide range of preferences over final outcomes, including linear preferences, converge to log preferences over each step. . . .

Eliezer Yudkowsky

> Eliezer, I think you are just mistaken; log preferences aggregate or split in time to log preferences.

Ah, okay, I see my problem. I was assuming that taking the log of population sizes just put us into a log-world, exchanging multiplication for addition. But in the new world, options add fixed amounts to your current total, regardless of your initial position, so preferences are just aggregative (not logarithmic) in the new world.

(*Thinks.*)

I think what this reveals is that, for repeatable choices with a certain kind of temporal independence and an indefinite time horizon, your local preferences will start corresponding to a representation under which the effect of those choices is purely aggregative, if such a representation exists. A representation where -4 units of negative is exactly balanced by $+1$ and $+3$ positive outcomes. As your time horizon approaches the indefinite, such an approach will dominate.

If you expect to encounter lots of options with nonmultiplicative effects—like "this will square my population, this will take the square root of my population"—then you'll be wise to regard those as $+1$ and -1 respectively, even though a logarithmic analysis will call this $+X$ vs. $-0.5X$.

Robin Hanson

Eliezer, it sounds like you are probably right with your ending comment, though it could be interesting to hear it elaborated, for a wider audience.

Eliezer Yudkowsky

Well, either you and I have really different visualizations of what the coherent parts of humanity's reflective equilibria would look like, or you don't think the Friendly AI project has the described outcome, or you have a really different moral reaction to that outcome.

If an AI goes FOOM, you seem to recognize that condition, or that prospect, as "total war." Afterward, you seem to recognize the resultant as a "God," and its relation to humanity as "rule." So either we've got really different visualizations of this process, or we have really different moral reactions to it. This seems worth exploring, because I suspect that it accounts for a large fraction of the real fuel in the argument.

Eliezer Yudkowsky

I don't consider myself a super-reliable math source. If the fate of the world isn't at stake, I'll often state an intuition rather than trying to prove it. For that matter, if the fate of the world *were* at stake, the first thing I'd do would be consult Marcello.

Robin, I accept the part about locally logarithmic behavior on spatially global and temporally local problems when there will be many future options and all are multiplicative. I don't accept the claim that evolution turns future entities into log-population maximizers. In a sense, you've actually shown just the opposite; *because* aggregative maximizers or log-maximizers will both show *instrumental* log-seeking behavior, entities with *terminal* log valuations have no fitness advantage. Evolution requires visible differences of behavior on which to operate.

If there are many nonmultiplicative options—say, there are ways to form trustworthy contracts, and a small party can contract with an intergalactic Warren Buffett—"I will give you 10% of my lineage's resources now, if you agree to use the same amount of resources to recreate copies of me in a billion years"—then it's not clear to me that logarithmics have an advantage; most of the numbers might be in aggregators because numbers are what they want, and that's what they use nonmultiplicative options to get.

Robin Hanson

Eliezer, I agree one might analyze nonmultiplicative worlds, but no one has done so yet, and the world so far has been pretty multiplicative. Please recall that I was initially responding to confident claims by Carl and others that evolution would make for terrible wars over the solar system because evolved creatures would be terminal-outcome-oriented and risk neutral about such outcomes. In this context I make three claims:

1. It is not obvious evolution would create terminal-outcome-oriented creatures.
2. It is not obvious such creatures would be risk-neutral about terminal outcomes.
3. Even if they were, they would have to be rather confident this conflict was in fact the last such conflict to be risk-neutral about resources gained from it.

Do you disagree with any of these claims?

Eliezer Yudkowsky

I don't know about *evolution* creating terminal-outcome-oriented creatures, but the case for self-modifying AIs by default converging to expected utility maximization has been written up by, e.g., Omohundro. But I think that what you mean here is aggregate valuation by expected utility maximizers. This wouldn't be *created* per se by either evolution or self-modification, but it also seems fairly likely to emerge as an idiom among utility functions not strictly specified. Other possible minds could be satisficers, and these would be less of a threat in a competitive situation (they would only take over the world if they knew they could win, or if they expected a strong threat to their button-to-keep-pressed if they weren't in sole charge of the galaxy).

Robin Hanson

I'm frustrated that I seem unable to communicate what should be a precise technical claim: evolution need *not* select for creatures who maximize expected future descendants. People keep claiming this as if it had been proven, but it has not, because it is not so.

The paper I cite is a clear precise counterexample. It considers a case where choices and probabilities are independent across time periods, and in this case it is optimal, *nonmyopically*, to make choices locally in time to max the expected log of period payoffs.

That case easily generalizes to chunks of N periods that are correlated arbitrarily internally, but independent across chunks. Again agents max the expected sum of log period returns, which is the same as maxing the expected sum of chunk returns. And you can make N as large as you like.

See original post for all comments.

* * *

1. Carl Shulman, "Zero and Non-zero-sum Games for Humans," private post, *Reflective Disequilibria* (blog), November 2008, http://reflectivedisequilibria.blogspot.com/2008/11/zero-and-nonzero-sum-games-for-humans.html.

2. Hans-Werner Sinn, "Weber's Law and the Biological Evolution of Risk Preferences: The Selective Dominance of the Logarithmic Utility Function," *Geneva Papers on Risk and Insurance Theory* 28, no. 2 (2003): 87–100, doi:10.1023/A:1026384519480.

3. Carl Shulman, "Evolutionary Selection of Preferences," private post, *Reflective Disequilibria* (blog), November 2008, http://reflectivedisequilibria.blogspot.com/2008/11/evolutionary-selection-of-preferences.html.

41

Sustained Strong Recursion

Eliezer Yudkowsky

5 December 2008

Followup to: Cascades, Cycles, Insight, Recursion, Magic

We seem to have a sticking point at the concept of "recursion," so I'll zoom in.

You have a friend who, even though he makes plenty of money, just spends all that money every month. You try to persuade your friend to *invest* a little—making valiant attempts to explain the wonders of compound interest by pointing to analogous processes in nature, like fission chain reactions.

"All right," says your friend, and buys a ten-year bond for $10,000, with an annual coupon of $500. Then he sits back, satisfied. "There!" he says. "Now I'll have an extra $500 to spend every year, without my needing to do any work! And when the bond comes due, I'll just roll

it over, so this can go on *indefinitely*. Surely, *now* I'm taking advantage of *the power of recursion*!"

"Um, no," you say. "That's not exactly what I had in mind when I talked about 'recursion.'"

"But I used some of my cumulative money earned to increase my very earning *rate*," your friend points out, quite logically. "If that's not 'recursion,' what *is?* My earning power has been 'folded in on itself,' just like you talked about!"

"Well," you say, "not exactly. Before, you were earning $100,000 per year, so your cumulative earnings went as $100{,}000 \cdot t$. Now, your cumulative earnings are going as $100{,}500 \cdot t$. That's not really much of a change. What we want is for your cumulative earnings to go as $B \cdot e^{A \cdot t}$ for some constants A and B—to grow *exponentially*."

"*Exponentially!*" says your friend, shocked.

"Yes," you say, "recursification has an amazing power to transform growth curves. In this case, it can turn a linear process into an exponential one. But to get that effect, you have to *reinvest the coupon payments* you get on your bonds—or at least reinvest some of them, instead of just spending them all. And you must be able to do this *over and over again*. Only *then* will you get the 'folding in' transformation, so that instead of your cumulative earnings going as $y = F(t) = A \cdot t$, your earnings will go as the differential equation $\frac{dy}{dt} = F(y) = A \cdot y$ whose solution is $y = e^{A \cdot t}$."

(I'm going to go ahead and leave out various constants of integration; feel free to add them back in.)

"Hold on," says your friend. "I don't understand the justification for what you just did there."

"Right now," you explain, "you're earning a steady income at your job, and you also have $500/year from the bond you bought. These are just things that go on generating money at a constant rate per unit time, in the background. So your cumulative earnings are the integral of that constant rate. If your earnings are y, then $\frac{dy}{dt} = A$, which resolves to $y = A \cdot t$. But now, suppose that, instead of having these constant earning forces operating in the background, we introduce a strong *feedback loop* from your cumulative earnings to your earning power."

"But I bought this one bond here—" says your friend.

"That's not enough for a *strong* feedback loop," you say. "Future increases in your cumulative earnings aren't going to increase the value of this one bond, or your salary, any *further*. One unit of force transmitted back is not a feedback loop—it has to be *repeatable*. You need a *sustained* recursion, not a one-off event."

"Okay," says your friend. "How about if I buy a $100 bond every year, then? Will *that* satisfy the strange requirements of this ritual?"

"Still not a strong feedback loop," you say. "Suppose that next year your salary went up $10,000/year—no, an even simpler example: suppose $10,000 fell in your lap out of the sky. If you only buy $100/year of bonds, that extra $10,000 isn't going to make any long-term difference to the earning curve. But if you're in the habit of investing 50% of found money, then there's a *strong* feedback loop from your cumulative earnings back to your earning power—we can pump up the cumulative earnings and watch the earning power rise as a direct result."

"How about if I just invest 0.1% of all my earnings, including the coupons on my bonds?" asks your friend.

"Well . . ." you say slowly. "That would be a *sustained* feedback loop but an extremely *weak* one, where marginal changes to your earnings have relatively small marginal effects on future earning power. I guess it would genuinely be a recursified process, but it would take a long time for the effects to become apparent, and any stronger recursions would easily outrun it."

"Okay," says your friend, "I'll start by investing a dollar, and I'll fully reinvest all the earnings from it, and the earnings on those earnings as well—"

"I'm not really sure there are any good investments that will let you invest just a dollar without it being eaten up in transaction costs," you say, "and it might not make a difference to anything on the timescales we have in mind—though there's an old story about a king, and grains of wheat placed on a chessboard . . . But realistically, a dollar isn't enough to get started."

"All right," says your friend, "suppose I start with $100,000 in bonds, and reinvest 80% of the coupons on those bonds plus rolling over all the principle, at a 5% interest rate, and we ignore inflation for now."

"Then," you reply, "we have the differential equation $\frac{dy}{dt} = 0.8 \cdot 0.05 \cdot y$, with the initial condition $y = \$100{,}000$ at $t = 0$, which works out to $y = \$100{,}000 \cdot e^{0.04 \cdot t}$. Or if you're reinvesting discretely rather than continuously, $y = \$100{,}000 \cdot (1.04)^t$."

We can similarly view the self-optimizing compiler in this light—it speeds itself up once, but never makes any further improvements, like buying a single bond; it's not a sustained recursion.

And now let us turn our attention to Moore's Law.

I am not a fan of Moore's Law. I think it's a red herring. I don't think you can forecast AI arrival times by using it, I don't think that AI (especially the good kind of AI) depends on Moore's Law continuing. I am agnostic about how long Moore's Law can continue—I simply leave the question to those better qualified, because it doesn't interest me very much . . .

But for our next simpler illustration of a strong recursification, we shall consider Moore's Law.

Tim Tyler serves us the duty of representing our strawman, repeatedly telling us, "But chip engineers use computers *now*, so Moore's Law is *already recursive*!"

To test this, we perform the equivalent of the thought experiment where we drop $10,000 out of the sky—push on the cumulative "wealth," and see what happens to the output rate.

Suppose that Intel's engineers could only work using computers of the sort available in 1998. How much would the next generation of computers be slowed down?

Suppose we gave Intel's engineers computers from 2018, in sealed black boxes (not transmitting any of 2018's knowledge). How much would Moore's Law speed up?

I don't work at Intel, so I can't actually answer those questions. I think, though, that if you said in the first case, "Moore's Law would drop way down, to something like 1998's level of improvement measured linearly in additional transistors per unit time," you would be way off base. And if you said in the second case, "I think Moore's Law would speed up by an order of magnitude, doubling every 1.8 months, until they caught up to the 2018 level," you would be equally way off base.

In both cases, I would expect the actual answer to be "not all that much happens." Seventeen instead of eighteen months, nineteen instead of eighteen months, something like that.

Yes, Intel's engineers have computers on their desks. But the serial speed or per-unit price of computing power is not, so far as I know, the limiting resource that bounds their research velocity. You'd probably have to ask someone at Intel to find out how much of their corporate income they spend on computing clusters/supercomputers, but I would guess it's not much compared to how much they spend on salaries or fab plants.

If anyone from Intel reads this, and wishes to explain to me how it would be unbelievably difficult to do their jobs using computers from ten years earlier, so that Moore's Law would slow to a crawl—then I stand ready to be corrected. But relative to my present state of partial knowledge, I would say that this does not look like a strong feedback loop.

However . . .

Suppose that the *researchers themselves* are running as uploads, software on the computer chips produced by their own factories.

Mind you, this is not the tiniest bit realistic. By my standards it's not even a very *interesting* way of looking at the Intelligence Explosion, because it does not deal with *smarter* minds but merely *faster* ones—it dodges the really difficult and interesting part of the problem.

Just as nine women cannot gestate a baby in one month; just as ten thousand researchers cannot do in one year what a hundred researchers can do in a hundred years; so too, a chimpanzee cannot do in four years what a human can do in one year, even though the chimp

has around one-fourth the human's cranial capacity. And likewise a chimp cannot do in a hundred years what a human does in ninety-five years, even though they share 95% of our genetic material.

Better-designed minds don't scale the same way as *larger* minds, and *larger* minds don't scale the same way as *faster* minds, any more than *faster* minds scale the same way as *more numerous* minds. So the notion of merely *faster* researchers, in my book, fails to address the interesting part of the "intelligence explosion."

Nonetheless, for the sake of illustrating this matter in a relatively simple case . . .

Suppose the researchers and engineers themselves—and the rest of the humans on the planet, providing a market for the chips and investment for the factories—are all running on the same computer chips that are the product of these selfsame factories. Suppose also that robotics technology stays on the same curve and provides these researchers with fast manipulators and fast sensors. We also suppose that the technology feeding Moore's Law has not yet hit physical limits. And that, as human brains are already highly parallel, we can speed them up even if Moore's Law is manifesting in increased parallelism instead of faster serial speeds—we suppose the uploads aren't *yet* being run on a fully parallelized machine, and so their actual serial speed goes up with Moore's Law. *Et cetera*.

In a fully naive fashion, we just take the economy the way it is today, and run it on the computer chips that the economy itself produces.

In our world where human brains run at constant speed (and eyes and hands work at constant speed), Moore's Law for computing power

s is:

$$s = R(t) = e^t.$$

The function R is the Research curve that relates the amount of Time t passed, to the current Speed of computers s.

To understand what happens when the researchers themselves are running on computers, we simply suppose that R does not relate computing technology to *sidereal* time—the orbits of the planets, the motion of the stars—but, rather, relates computing technology to the amount of subjective time spent researching it.

Since in *our* world subjective time is a linear function of sidereal time, this hypothesis fits *exactly the same curve* R to observed human history so far.

Our direct measurements of observables do not constrain between the two hypotheses:

1. Moore's Law is exponential in the number of orbits of Mars around the Sun.

2. Moore's Law is exponential in the amount of subjective time that researchers spend thinking and experimenting and building using a proportional amount of sensorimotor bandwidth.

But our prior knowledge of causality may lead us to prefer the second hypothesis.

So to understand what happens when the Intel engineers themselves run on computers (and use robotics) subject to Moore's Law, we recursify and get:

$$\frac{dy}{dt} = s = R(y) = e^y.$$

Here y is the total amount of elapsed *subjective* time, which at any given point is increasing according to the computer speed s given by Moore's Law, which is determined by the same function R that describes how Research converts elapsed subjective time into faster computers. Observed human history to date roughly matches the hypothesis that R is exponential with a doubling time of eighteen subjective months (or whatever). Solving

$$\frac{dy}{dt} = e^y$$

yields

$$y = -\ln(C - t).$$

One observes that this function goes to +infinity at a finite time C.

This is only to be expected, given our assumptions. After eighteen sidereal months, computing speeds double; after another eighteen subjective months, or nine sidereal months, computing speeds double again; etc.

Now, unless the physical universe works in a way that is not only *different* from the current standard model, but has a different *character of physical law* than the current standard model; you can't *actually* do infinite computation in finite time.

Let us suppose that if our biological world had no Intelligence Explosion, and Intel just kept on running as a company, populated by humans, forever, that Moore's Law would start to run into trouble around 2020. Say, after 2020 there would be a ten-year gap where chips simply stagnated, until the next doubling occurred after a hard-won breakthrough in 2030.

This just says that $R(y)$ is not an indefinite exponential curve. By hypothesis, from subjective years 2020 to 2030, $R(y)$ is flat, corresponding to a constant computer speed s. So $\frac{dy}{dt}$ is constant over this same time period: Total elapsed subjective time y grows at a linear rate, and as y grows, $R(y)$ and computing speeds remain flat until ten subjective years have passed. So the *sidereal* bottleneck lasts ten subjective years times the current sidereal/subjective conversion rate at 2020's computing speeds.

In short, the whole scenario behaves exactly like what you would expect—the simple transform really does describe the naive scenario of "drop the economy into the timescale of its own computers."

After subjective year 2030, things pick up again, maybe—there are ultimate physical limits on computation, but they're pretty damned high, and we've got a ways to go until there. But maybe Moore's Law is slowing down—going subexponential, and then, as the physical limits are approached, logarithmic, and then simply giving out.

But whatever your beliefs about where Moore's Law ultimately goes, you can just map out the way you would expect the research function R to work as a function of sidereal time in our own world, and then apply the transformation $\frac{dy}{dt} = R(y)$ to get the progress of the uploaded civilization over sidereal time t. (Its progress over *subjective* time is simply given by R.)

If sensorimotor bandwidth is the critical limiting resource, then we instead care about R&D on fast sensors and fast manipulators. We want $R_{sm}(y)$ instead $R(y)$, where R_{sm} is the progress rate of sensors and manipulators as a function of elapsed sensorimotor time. And then we write $\frac{dy}{dt} = R_{sm}(y)$ and crank on the equation again to find out what the world looks like from a sidereal perspective.

We can verify that the Moore's Researchers scenario is a strong positive feedback loop by performing the "drop $10,000" thought experiment. Say, we drop in chips from another six doublings down the road—letting the researchers run on those faster chips, while holding constant their state of technological knowledge.

Lo and behold, this drop has a rather *large* impact, much larger than the impact of giving faster computers to our own biological world's Intel. *Subjectively* the impact may be unnoticeable—as a citizen, you just see the planets slow down again in the sky. But sidereal growth rates increase by a factor of sixty-four.

So this is indeed deserving of the names "strong positive feedback loop" and "sustained recursion."

As disclaimed before, all this isn't *really* going to happen. There would be effects like those Robin Hanson prefers to analyze, from being able to spawn new researchers as the cost of computing power decreased. You might be able to pay more to get researchers twice as fast. Above all, someone's bound to try hacking the uploads for increased intelligence . . . and then those uploads will hack themselves even further . . . Not to mention that it's not clear how this civilization cleanly dropped into computer time in the first place.

So no, this is not supposed to be a realistic vision of the future.

But, alongside our earlier parable of compound interest, it *is* supposed to be an illustration of how strong, sustained recursion has much more drastic effects on the shape of a growth curve than a one-off case of one thing leading to another thing. Intel's engineers *running on* computers is not like Intel's engineers *using* computers.

* * *

Robin Hanson

You can define "recursive" as accelerating growth, in which case it remains an open question whether any particular scenario, such as sped-up folks researching how to speed up, is in fact recursive. Or you can, as I had thought you did, define "recursive" as a situation of a loop of growth factors each encouraging the next one in the loop, in which case it is an open question if that results in accelerating growth. I was pointing out before that there exist loops of encouraging growth factors that do not result in accelerating growth. If you choose the other definition strategy, I'll note that your model is extremely stark and leaves out the usual items in even the simplest standard growth models.

Eliezer Yudkowsky

Robin, like I say, most AIs won't hockey-stick, and when you fold a function in on itself this way, it can bottleneck for a billion years if its current output is flat or bounded. That's why self-optimizing compilers don't go FOOM.

"Recursion" is not accelerating growth. It is not a loop of growth factors. "Adding a recursion" describes situations where you might naively be tempted to take an existing function

$$y = F(t)$$

and rewrite it as

$$\frac{dy}{dt} = F(y).$$

Does that make it any clearer?

Robin Hanson

Eliezer, if "adding a recursion" means adding one more power to the derivative in the growth equation, then it is an open question what sorts of AIs would do that. And then it isn't clear why you would say Engelbart was "not re-

cursive enough," since this is a discrete definition without some parameter you can have not enough of.

Eliezer Yudkowsky

Robin, how is the transition

$$y = e^t \Rightarrow \frac{dy}{dt} = e^t$$

to

$$\frac{dy}{dt} = e^y \Rightarrow y = -\ln(C - t) \Rightarrow \frac{dy}{dt} = \frac{1}{C - t}$$

"adding one more power to the derivative in the growth equation"?

I'm not sure what that phrase you used means, exactly, but I wonder if you may be mis-visualizing the general effect of what I call "recursion."

Or what about

$$y = t^2 \qquad \rightarrow \qquad \frac{dy}{dt} = y^2$$

etc. Or

$$y = \log t \qquad \rightarrow \qquad \frac{dy}{dt} = \log y,$$

etc.

Like I said, this doesn't necessarily hockey-stick; if you get sublinear returns the recursified version will be slower than the original.

Eliezer Yudkowsky

Engelbart was "not recursive enough" in the sense that he didn't have a *strong, sustained* recursion; his tech improvements did not yield an increase in

engineering velocity which was sufficient to produce tech improvements that would further improve his engineering velocity. He wasn't running on his own chips. Like EURISKO, he used his scientific prowess to buy some bonds (computer tech) that paid a relatively low coupon on further scientific prowess, and the interest payments didn't let him buy all that many more bonds.

Robin Hanson

In the post and comment discussion with me Eliezer tries to offer a math definition of "recursive" but in this discussion about Intel he seems to revert to the definition I thought he was using all along, about whether growing X helps Y grow better which helps X grow better. I don't see any differential equations in the Intel discussion.

Eliezer Yudkowsky

Does it help if I say that "recursion" is not something which is true or false of a given system, but rather something by which one version of a system *differs* from another?

The question is not "Is Intel recursive?" but rather "Which of these two systems is the case? Does intervening on Intel to provide them with much less or much more computing power tremendously slow or accelerate their progress? Or would it have only small fractional effects?"

In the former case, the research going into Moore's Law is being kept *rigidly* on track by the computers' output by Moore's Law, and this would make it plausible that the exponential form of Moore's Law was due *primarily* to this effect.

In the latter case, computing power is only loosely coupled to Intel's research activities, and we have to search for other explanations for Moore's Law, such as that the market's sensitivity to computing power is logarithmic and so Intel scales its resources as high as necessary to achieve a certain multiplicative improvement, but no higher than that. . . .

Robin Hanson

Eliezer, I don't know what is your implicit referent to divide "tremendous" from "fractional" influence of growth of X on growth of Y. Perhaps you can define that clearly in a very simple model, but I don't see how to generalize that to more realistic models. . . .

See original post for all comments.

42

Friendly Projects vs. Products

Robin Hanson

5 December 2008

I'm a big board game fan, and my favorite these days is *Imperial*. *Imperial* looks superficially like the classic strategy-intense war game *Diplomacy*, but with a crucial difference: instead of playing a nation trying to win WWI, you play a banker trying to make money from that situation. If a nation you control (by having loaned it the most) is threatened by another nation, you might indeed fight a war, but you might instead just buy control of that nation. This is a great way to mute conflicts in a modern economy: have conflicting groups buy shares in each other.

For projects to create new creatures, such as ems or AIs, there are two distinct friendliness issues:

Project Friendliness: *Will the race make winners and losers, and how will winners treat losers?* While any race might be treated as part of a

total war on several sides, usually the inequality created by the race is moderate and tolerable. For larger inequalities, projects can explicitly join together, agree to cooperate in weaker ways such as by sharing information, or they can buy shares in each other. Naturally arising info leaks and shared standards may also reduce inequality even without intentional cooperation. The main reason for failure here would seem to be the sorts of distrust that plague all human cooperation.

Product Friendliness: *Will the creatures cooperate with or rebel against their creators?* Folks running a project have reasonably strong incentives to avoid this problem. Of course for the case of extremely destructive creatures the project might internalize more of the gains from cooperative creatures than they do the losses from rebellious creatures. So there might be some grounds for wider regulation. But the main reason for failure here would seem to be poor judgment, thinking you had your creatures more surely under control than in fact you did.

It hasn't been that clear to me which of these is the main concern re "friendly AI."

Added: Since Eliezer says product friendliness is his main concern, let me note that the main problem there is the tails of the distribution of *bias* among project leaders. If all projects agreed the problem was very serious they would take near-appropriate caution to isolate their creatures, test creature values, and slow creature development enough to track progress sufficiently. Designing and

advertising a solution is one approach to reducing this bias, but it need not need the best approach; perhaps institutions like prediction markets that aggregate info and congeal a believable consensus would be more effective.

* * *

Eliezer Yudkowsky

The second one, he said without the tiniest trace of hesitation.

Robin Hanson

I just added to the post.

Eliezer Yudkowsky

> If all projects agreed the problem was very serious they would take near-appropriate caution to isolate their creatures, test creature values, and slow creature development enough to track progress sufficiently.

Robin, I agree this is a left-tail problem, or to be more accurate, the right tail of the left hump of a two-hump camel.

But your suggested description of a solution *is not going to work*. You need something that can carry out a billion sequential self-modifications on itself without altering its terminal values, and you need exactly the right terminal values because missing or distorting a single one can spell the difference between utopia or dystopia. The former requires new math, the latter requires extremely meta thinking plus additional new math. *If no one has this math, all good guys are helpless* and the game is lost automatically.

That's why I see this as currently having the status of a math problem even more than a PR problem.

For all the good intentions that ooze from my every pore, right now I do not, technically speaking, *know* how to build a Friendly AI—though thankfully, I know enough to know why "testing" isn't a solution (context not i.i.d.) which removes me from the right tail of the left hump.

Now, some aspects of this can be viewed as a PR problem—you want to remove researchers from the right tail of the left hump, which you can do up to a point through publicizing dangers. And you want to add researchers to the right tail of the right hump, which you can do by, among other strategies, having math geniuses read *Overcoming Bias* at age fifteen and then waiting a bit. (Some preliminary evidence indicates that this strategy may already be working.)

But above all, humanity is faced with a win-or-fail *math* problem, a challenge of pure technical knowledge stripped of all social aspects. It's not that this is the only part of the problem. It's just the only impossible part of the problem.

Robin Hanson

. . . Eliezer, I'd like to hear more about why testing and monitoring creatures as they develop through near-human levels, slowing development as needed, says nothing useful about their values as transhuman creatures. And about why it isn't enough to convince most others that the problem is as hard as you say: in that case many others would also work to solve the problem, and would avoid inducing it until they had a solution. And hey, if you engage them there's always a chance they'll convince you they are right and you are wrong. Note that your social strategy, of avoiding standard credentials, is about the worst case for convincing a wide audience.

See original post for all comments.

43

Is That Your True Rejection?

Eliezer Yudkowsky

6 December 2008

It happens every now and then that the one encounters some of my transhumanist-side beliefs—as opposed to my ideas having to do with human rationality—strange, exotic-sounding ideas like superintelligence and Friendly AI. And the one rejects them.

If the one is called upon to explain the rejection, not uncommonly the one says,

"Why should I believe anything Yudkowsky says? He doesn't have a PhD!"

And occasionally someone else, hearing, says, "Oh, you should get a PhD, so that people will listen to you." Or this advice may even be offered by the same one who disbelieved, saying, "Come back when you have a PhD."

Now there are good and bad reasons to get a PhD, but this is one of the bad ones.

There's many reasons why someone *actually* has an adverse reaction to transhumanist theses. Most are matters of pattern recognition, rather than verbal thought: the thesis matches against "strange weird idea" or "science fiction" or "end-of-the-world cult" or "overenthusiastic youth."

So immediately, at the speed of perception, the idea is rejected. If, afterward, someone says, "Why not?" this launches a search for justification. But this search will not necessarily hit on the true reason—by "true reason" I mean not the *best* reason that could be offered, but rather, whichever causes were decisive as a matter of historical fact, at the *very first* moment the rejection occurred.

Instead, the search for justification hits on the justifying-sounding fact, "This speaker does not have a PhD."

But I also don't have a PhD when I talk about human rationality, so why is the same objection not raised there?

And more to the point, if I *had* a PhD, people would not treat this as a decisive factor indicating that they ought to believe everything I say. Rather, the same initial rejection would occur, for the same reasons; and the search for justification, afterward, would terminate at a different stopping point.

They would say, "Why should I believe *you?* You're just some guy with a PhD! There are lots of those. Come back when you're well-known in your field and tenured at a major university."

But do people *actually* believe arbitrary professors at Harvard who say weird things? Of course not. (But if I were a professor at Harvard, it would in fact be easier to get *media attention*. Reporters initially dis-

inclined to believe me—who would probably be equally disinclined to believe a random PhD-bearer—would still report on me, because it would be news that a Harvard professor believes such a weird thing.)

If you are saying things that sound *wrong* to a novice, as opposed to just rattling off magical-sounding technobabble about leptical quark braids in $N+2$ dimensions; and the hearer is a stranger, unfamiliar with you personally *and* with the subject matter of your field; then I suspect that the point at which the average person will *actually* start to grant credence overriding their initial impression, purely *because* of academic credentials, is somewhere around the Nobel Laureate level. If that. Roughly, you need whatever level of academic credential qualifies as "beyond the mundane."

This is more or less what happened to Eric Drexler, as far as I can tell. He presented his vision of nanotechnology, and people said, "Where are the technical details?" or, "Come back when you have a PhD!" And Eric Drexler spent six years writing up technical details and got his PhD under Marvin Minsky for doing it. And *Nanosystems* is a great book. But did the same people who said, "Come back when you have a PhD," actually change their minds at all about molecular nanotechnology? Not so far as I ever heard.

It has similarly been a general rule with the Machine Intelligence Research Institute that, whatever it is we're supposed to do to be more credible, when we actually do it, nothing much changes. "Do you do any sort of code development? I'm not interested in supporting an organization that doesn't develop code" → OpenCog → nothing changes. "Eliezer Yudkowsky lacks academic credentials" → Professor Ben Goertzel installed as Director of Research → nothing changes. The one thing that actually *has* seemed to raise credibility

is famous people associating with the organization, like Peter Thiel funding us, or Ray Kurzweil on the Board.

This might be an important thing for young businesses and new-minted consultants to keep in mind—that what your failed prospects *tell* you is the reason for rejection may not make the *real* difference, and you should ponder that carefully before spending huge efforts. If the venture capitalist says, "If only your sales were growing a little faster!"—if the potential customer says, "It seems good, but you don't have feature X"—that may not be the true rejection. Fixing it may or may not change anything.

And it would also be something to keep in mind during disagreements. Robin and I share a belief that two rationalists should not agree to disagree: they should not have common knowledge of epistemic disagreement unless something is very wrong.

I suspect that, in general, if two rationalists set out to resolve a disagreement that persisted past the first exchange, they should expect to find that the true sources of the disagreement are either hard to communicate, or hard to expose. E.g:

- Uncommon, but well-supported, scientific knowledge or math
- Long inferential distances
- Hard-to-verbalize intuitions, perhaps stemming from specific visualizations
- Zeitgeists inherited from a profession (which may have good reason for it)
- Patterns perceptually recognized from experience

- Sheer habits of thought
- Emotional commitments to believing in a particular outcome
- Fear of a past mistake being disproven
- Deep self-deception for the sake of pride or other personal benefits

If the matter were one in which *all* the true rejections could be *easily* laid on the table, the disagreement would probably be so straightforward to resolve that it would never have lasted past the first meeting.

"Is this my true rejection?" is something that both disagreers should surely be asking *themselves*, to make things easier on the Other Fellow. However, attempts to directly, publicly psychoanalyze the Other may cause the conversation to degenerate *very* fast, in my observation.

Still—"Is that your true rejection?" should be fair game for Disagreers to humbly ask, if there's any productive way to pursue that subissue. Maybe the rule could be that you can openly ask, "Is that simple straightforward-sounding reason your *true* rejection, or does it come from intuition X or professional zeitgeist Y?" While the more embarrassing possibilities lower on the table are left to the Other's conscience, as their own responsibility to handle.

***Post scriptum*:** This post is not *really* about PhDs in general, or their credibility value in particular. But I've always figured that, to the extent this was a strategically important consideration, it would make more sense to recruit an academic of existing high status than

spend a huge amount of time trying to achieve low or moderate academic status.

However, if any professor out there wants to let me come in and *just* do a PhD in analytic philosophy—*just* write the thesis and defend it—then I have, for my own use, worked out a general and mathematically elegant theory of Newcomb-like decision problems. I think it would make a fine PhD thesis, and it is ready to be written—if anyone has the power to let me do things the old-fashioned way.

* * *

Robin Hanson

There need not be just one "true objection"; there can be many factors that together lead to an estimate. Whether you have a PhD, and whether folks with PhDs have reviewed your claims, and what they say, can certainly be relevant. Also remember that you should care lots more about the opinions of experts that could build on and endorse your work than about average-Joe opinions. Very few things ever convince average folks of anything unusual; target a narrower audience.

Eliezer Yudkowsky

. . . Robin, see the *post scriptum*. I would be willing to get a PhD thesis if it went by the old rules and the old meaning of "Prove you can make an original, significant contribution to human knowledge and that you've mastered an existing field," rather than "This credential shows you have spent X number of years in a building." (This particular theory *would* be hard enough to write up that I may not get around to it if a PhD credential isn't at stake.)

See original post for all comments.

44

Shared AI Wins

Robin Hanson

6 December 2008

Almost every new technology comes at first in a dizzying variety of styles and then converges to what later seems the "obvious" configuration. It is actually quite an eye-opener to go back and see old might-have-beens, from steam-powered cars to pneumatic tube mail to memex to Engelbart's computer tools. Techs that are only imagined, not implemented, take on the widest range of variations. When actual implementations appear, people slowly figure out what works better, while network and other scale effects lock in popular approaches. As standards congeal, competitors focus on smaller variations around accepted approaches. Those who stick with odd standards tend to be marginalized.

Eliezer says standards barriers are why AIs would "foom" locally, with one AI quickly growing from so small no one notices to so powerful it takes over the world:

> I also don't think this [scenario] is allowed: . . . knowledge and even skills are widely traded in this economy of AI systems. In concert, these AIs, and their human owners, and the economy that surrounds them, undergo a *collective* FOOM of self-improvement. No local agent is capable of doing all this work, only the collective system. . . .
>
> [The reason is that] trading cognitive content around between diverse AIs is more difficult and less likely than it might sound. Consider the field of AI as it works today. Is there *any* standard database of cognitive content that you buy off the shelf and plug into your amazing new system, whether it be a chess player or a new data-mining algorithm? . . .
>
> . . . The diversity of cognitive architectures acts as a *tremendous* barrier to trading around cognitive content. . . . If two AIs both see an apple for the first time, and they both independently form concepts about that apple . . . their *thoughts* are effectively written in a different language. . . .
>
> The barrier this opposes to a true, cross-agent, literal "economy of mind," is so strong, that in the vast majority of AI applications you set out to write today, you will not bother to import any standardized preprocessed cognitive content. It will be easier for your AI application to start with some standard examples—databases of *that* sort of thing do exist, in some fields anyway—and *redo all the cognitive work of learning* on its own. . . .
>
> . . . Looking over the diversity of architectures proposed at any AGI conference I've attended, it is very hard to imag-

> ine directly trading cognitive content between any two of them.

But *of course* "visionaries" take a wide range of incompatible approaches. Commercial software tries much harder to match standards and share sources. The whole point of Cyc was that AI researchers neglect compatibility and sharing because they are more interested in writing papers than making real systems. The idea that you could create human-level intelligence by just feeding raw data into the right math-inspired architecture is pure fantasy. You couldn't build an effective cell or ecosystem or developed economy or most any complex system that way either—such things require not just good structure but also lots of good content. Loners who start all over from scratch rarely beat established groups sharing enough standards to let them share improvements to slowly accumulate content.

Cyc content may or may not jump-start a sharing AI community, but AI just won't happen without a whole lot of content. If ems appear first, perhaps shareable em contents could form a different basis for shared improvements.

* * *

Eliezer Yudkowsky

It's generally a terrible analogy, but would you say that a human baby growing up is getting "raw data" fed into the right architecture, or that human babies are exposed to data preprocessed by their parents, or that human babies get standardized data?

Robin Hanson

. . . Eliezer, a human baby certainly gets raw data, and it has a good architecture too, but in addition I'd say it has lots of genetically encoded info about what sort of patterns in data to expect and attend to, i.e., what sort of abstractions to consider. In addition, when raising kids we focus their attention on relevant and useful patterns and abstractions. And of course we just tell them lots of stuff too. . . .

Eliezer Yudkowsky

This is much like my visualization of how an AI works, except that there's substantially less "genetically encoded info" at the time you boot up the system—mostly consisting of priors that have to be encoded procedurally. This is work done by natural selection in the case of humans; so some of that is taken off your hands by programs that you write, and some of it is work you do at runtime over the course of the AI's development, rather than trying to encode into the very first initial system. But you can't exactly leave out Bayes' Rule, or causal graphs, or *modus ponens*, from the first system. . . .

Robin Hanson

. . . Eliezer, yes, well-chosen priors *are* the key "encoded info." There may be a misunderstanding that when I say "info" people think I mean direct facts like "Paris is capital of France," while I instead mean any content within your architecture that helps you focus attention well. Clearly human babies do leave out Bayes' Rule and *modus ponens*, but yes, we should put that in if we can cleanly do so. I'd just claim that doesn't get you very far; you'll need to find a way to inherit big chunks of the vast human content heritage.

Eliezer Yudkowsky

Robin, "Bayes' Rule" doesn't mean a little declarative representation of Bayes' Rule, it means updating in response to evidence that seems more likely in one case than another. Hence "encoded procedurally."

Robin Hanson

Eliezer, yes, babies clearly do approximately encode some implications of Bayes' Rule, but also clearly fail to encode many other implications.

See original post for all comments.

45

Artificial Mysterious Intelligence

Eliezer Yudkowsky

7 December 2008

Previously in series: Failure By Affective Analogy

I once had a conversation that I still remember for its sheer, purified archetypicality. This was a nontechnical guy, but pieces of this dialog have also appeared in conversations I've had with professional AI folk . . .

> Him: Oh, you're working on AI! Are you using neural networks?
>
> Me: I think emphatically *not*.
>
> Him: But neural networks are so wonderful! They solve problems and we don't have any idea how they do it!
>
> Me: If you are ignorant of a phenomenon, that is a fact about your state of mind, not a fact about the phenomenon

itself. Therefore your ignorance of how neural networks are solving a specific problem cannot be responsible for making them work better.

Him: Huh?

Me: If you don't know how your AI works, that is not good. It is bad.

Him: Well, intelligence is much too difficult for us to understand, so we need to find *some* way to build AI without understanding how it works.

Me: Look, even if you could do that, you wouldn't be able to predict any kind of positive outcome from it. For all you knew, the AI would go out and slaughter orphans.

Him: Maybe we'll build Artificial Intelligence by scanning the brain and building a neuron-by-neuron duplicate. Humans are the only systems we know are intelligent.

Me: It's hard to build a flying machine if the only thing you understand about flight is that somehow birds magically fly. What you need is a concept of aerodynamic lift, so that you can see how something can fly even if it isn't exactly like a bird.

Him: That's too hard. We have to copy something that we know works.

Me: (*reflectively*) What do people find so unbearably *awful* about the prospect of having to finally break down and solve the bloody problem? Is it really *that* horrible?

Him: Wait . . . you're saying you want to actually *understand* intelligence?

Me: Yeah.

Him: (*aghast*) Seriously?

Me: I don't know everything I need to know about intelligence, but I've learned a hell of a lot. Enough to know what happens if I try to build AI while there are still gaps in my understanding.

> Him: Understanding the problem is too hard. You'll never do it.

That's not just a difference of opinion you're looking at, it's a *clash of cultures*.

For a long time, many different parties and factions in AI, adherent to more than one ideology, have been trying to build AI *without* understanding intelligence. And their habits of thought have become ingrained in the field, and even transmitted to parts of the general public.

You may have heard proposals for building true AI which go something like this:

1. Calculate how many operations the human brain performs every second. This is "the only amount of computing power that we know is actually sufficient for human-equivalent intelligence." Raise enough venture capital to buy a supercomputer that performs an equivalent number of floating-point operations in one second. Use it to run the most advanced available neural network algorithms.

2. The brain is huge and complex. When the Internet becomes sufficiently huge and complex, intelligence is bound to emerge from the Internet. *(I get asked about this in 50% of my interviews.)*

3. Computers seem unintelligent because they lack common sense. Program a very large number of "common-sense facts" into a computer. Let it try to reason about the relation of these

facts. Put a sufficiently huge quantity of knowledge into the machine, and intelligence will emerge from it.

4. Neuroscience continues to advance at a steady rate. Eventually, super-MRI or brain sectioning and scanning will give us precise knowledge of the local characteristics of all human brain areas. So we'll be able to build a duplicate of the human brain by duplicating the parts. "The human brain is the only example we have of intelligence."

5. Natural selection produced the human brain. It is "the only method that we know works for producing general intelligence." So we'll have to scrape up a really huge amount of computing power, and *evolve* AI.

What do all these proposals have in common?

They are all ways to make yourself believe that you can build an Artificial Intelligence even if you don't understand exactly how intelligence works.

Now, such a belief is not necessarily *false*! Methods (4) and (5), if pursued long enough and with enough resources, *will* eventually work. (Method (5) might require a computer the size of the Moon, but give it *enough* crunch and it will work, even if you have to simulate a quintillion planets and not just one . . .)

But regardless of whether any given method would work in principle, the unfortunate habits of thought will already begin to arise as soon as you start thinking of ways to create Artificial Intelligence without having to penetrate the *mystery of intelligence*.

I have already spoken of some of the hope-generating tricks that appear in the examples above. There is invoking similarity to humans, or using words that make you feel good. But really, a lot of the trick here just consists of imagining yourself hitting the AI problem with a *really big rock*.

I know someone who goes around insisting that AI will cost a quadrillion dollars, and as soon as we're willing to spend a quadrillion dollars, we'll have AI, and we couldn't possibly get AI without spending a quadrillion dollars. "Quadrillion dollars" is his big rock that he imagines hitting the problem with, even though he doesn't quite understand it.

It often will not occur to people that the mystery of intelligence could be any more penetrable than it *seems*: By the power of the Mind Projection Fallacy, being ignorant of how intelligence works will make it seem like intelligence is inherently impenetrable and chaotic. They will think they possess a positive knowledge of intractability, rather than thinking, "I am ignorant."

And the thing to remember is that, for these last decades on end, *any* professional in the field of AI trying to build "real AI" had some reason for trying to do it without really understanding intelligence (various fake reductions aside).

The New Connectionists accused the Good Old-Fashioned AI researchers of not being parallel enough, not being fuzzy enough, not being emergent enough. But they did not say, "There is too much you do not understand."

The New Connectionists catalogued the flaws of GOFAI for years on end, with fiery castigation. But they couldn't ever actually say: "How *exactly* are all these logical deductions going to produce 'intel-

ligence,' anyway? Can you walk me through the cognitive operations, step by step, which lead to that result? Can you explain 'intelligence' and how you plan to get it, without pointing to humans as an example?"

For they themselves would be subject to exactly the same criticism.

In the house of glass, somehow, no one ever gets around to talking about throwing stones.

To tell a lie, you have to lie about all the other facts entangled with that fact, and also lie about the methods used to arrive at beliefs: The culture of Artificial Mysterious Intelligence has developed its own Dark Side Epistemology, complete with reasons why it's actually *wrong* to try and understand intelligence.

Yet when you step back from the bustle of this moment's history, and think about the long sweep of science—there was a time when stars were mysterious, when chemistry was mysterious, when life was mysterious. And in this era, much was attributed to black-box essences. And there were many hopes based on the similarity of one thing to another. To many, I'm sure, alchemy just seemed very *difficult* rather than even seeming *mysterious*; most alchemists probably did not go around thinking, "Look at how much I am disadvantaged by not knowing about the existence of chemistry! I must discover atoms and molecules as soon as possible!" They just memorized libraries of random things you could do with acid and bemoaned how difficult it was to create the Philosopher's Stone.

In the end, though, what happened is that scientists achieved insight, and *then* things got much easier to do. You also had a better

idea of what you could or couldn't do. The problem stopped being *scary* and *confusing*.

But you wouldn't hear a New Connectionist say, "Hey, maybe all the failed promises of 'logical AI' were basically due to the fact that, in their epistemic condition, they had no right to expect their AIs to work in the first place, because they couldn't actually have sketched out the link in any more detail than a medieval alchemist trying to explain why a particular formula for the Philosopher's Stone will yield gold." It would be like the Pope attacking Islam on the basis that faith is not an adequate justification for asserting the existence of their deity.

Yet, in fact, the promises *did* fail, and so we can conclude that the promisers overreached what they had a right to expect. The Way is not omnipotent, and a bounded rationalist cannot do all things. But even a bounded rationalist can aspire not to overpromise—to only *say* you can do that which you *can* do. So if we want to achieve that reliably, history shows that we should not accept certain kinds of hope. In the absence of insight, hopes tend to be unjustified because you lack the knowledge that would be needed to justify them.

We humans have a difficult time working in the absence of insight. It doesn't reduce us all the way down to being as stupid as evolution. But it makes everything difficult and tedious and annoying.

If the prospect of having to finally break down and solve the bloody problem of intelligence seems scary, you underestimate the interminable hell of *not* solving it.

* * *

Robin Hanson

We shouldn't underrate the power of insight, but we shouldn't overrate it either; some systems can just be a mass of details, and to master such systems you must master those details. And if you pin your hopes for AI progress on powerful future insights, you have to ask how often such insights occur, and how many we would need. The track record so far doesn't look especially encouraging.

Eliezer Yudkowsky

Robin, the question of whether compact insights *exist* and whether they are *likely to be obtained in reasonable time* (and by how large a group, etc.) are very different questions and should be considered separately, in order. . . .

See original post for all comments.

46

Wrapping Up

Robin Hanson

7 December 2008

This Friendly AI discussion has taken more time than I planned or have. So let me start to wrap up.

On small scales we humans evolved to cooperate via various pair and group bonding mechanisms. But these mechanisms aren't of much use on today's evolutionarily unprecedented large scales. Yet we do in fact cooperate on the largest scales. We do this because we are risk averse, because our values mainly conflict on resource use which conflicts destroy, and because we have the intelligence and institutions to enforce win-win deals via property rights, etc.

I raise my kids because they share my values. I teach other kids because I'm paid to. Folks raise horses because others pay them for horses, expecting horses to cooperate as slaves. You might expect

your pit bulls to cooperate, but we should only let you raise pit bulls if you can pay enough damages if they hurt your neighbors.

In my preferred em (whole-brain emulation) scenario, people would only authorize making em copies using borrowed or rented brains/bodies when they expected those copies to have lives worth living. With property rights enforced, both sides would expect to benefit more when copying was allowed. Ems would not exterminate humans mainly because that would threaten the institutions ems use to keep peace with each other.

Similarly, we expect AI developers to plan to benefit from AI cooperation via either direct control, indirect control such as via property-rights institutions, or such creatures having cooperative values. As with pit bulls, developers should have to show an ability, perhaps via insurance, to pay plausible hurt amounts if their creations hurt others. To the extent they or their insurers fear such hurt, they would test for various hurt scenarios, slowing development as needed in support. To the extent they feared inequality from some developers succeeding first, they could exchange shares, or share certain kinds of info. Naturally occurring info leaks, and shared sources, both encouraged by shared standards, would limit this inequality.

In this context, I read Eliezer as fearing that developers, insurers, regulators, and judges will vastly underestimate how dangerous are newly developed AIs. *Eliezer guesses that within a few weeks a single AI could grow via largely internal means from weak and unnoticed to so strong it takes over the world,* with no weak but visible moment between when others might just nuke it. Since its growth needs little from the rest of the world, and since its resulting power is so vast, only its values would make it treat others as much more than raw materi-

als. But its values as seen when weak say little about its values when strong. Thus Eliezer sees little choice but to try to design a theoretically clean AI architecture allowing near-provably predictable values when strong, to in addition design a set of robust good values, and then to get AI developers to adopt this architecture/values combination.

This is not a choice to make lightly; declaring your plan to build an AI to take over the world would surely be seen as an act of war by most who thought you could succeed, no matter how benevolent you said its values would be. (But yes, if Eliezer were sure, he should push ahead anyway.) And note most of Eliezer's claim's urgency comes from the fact that most of the world, including most AI researchers, *disagree* with Eliezer; if they agreed, AI development would likely be severely regulated, like nukes today.

On the margin this scenario seems less a concern when manufacturing is less local, when tech surveillance is stronger, and when intelligence is multidimensional. It also seems less of a concern with ems, as AIs would have less of a hardware advantage over ems, and modeling AI architectures on em architectures would allow more reliable value matches.

While historical trends do suggest we watch for a several-year-long transition sometime in the next century to a global growth rate two or three orders of magnitude faster, Eliezer's postulated local growth rate seems much faster. I also find Eliezer's growth math unpersuasive. Usually dozens of relevant factors are coevolving, with several loops of, all else equal, X growth speeds Y growth speeds etc. Yet usually it all adds up to exponential growth, with rare jumps to faster growth rates. Sure, if you pick two things that plausibly speed

each other and leave everything else out including diminishing returns, your math can suggest accelerating growth to infinity, but for a real foom that loop needs to be real strong, much stronger than contrary muting effects.

But the real sticking point seems to be locality. The "content" of a system is its small modular features while its "architecture" is its most important, least modular features. Imagine a large community of AI developers, with real customers, mostly adhering to common architectural standards and sharing common content; imagine developers trying to gain more market share and that AIs mostly got better by accumulating more better content, and that this rate of accumulation mostly depended on previous content; imagine architecture is a minor influence. In this case the whole AI sector of the economy might grow very quickly, but it gets pretty hard to imagine one AI project zooming vastly ahead of others.

So I suspect this all comes down to, how powerful is architecture in AI, and how many architectural insights can be found how quickly? If there were say a series of twenty deep powerful insights, each of which made a system twice as effective, just enough extra oomph to let the project and system find the next insight, it would add up to a factor of a million. Which would still be nowhere near enough, so imagine a lot more of them, or lots more powerful.

This scenario seems quite flattering to Einstein wannabes, making deep-insight-producing Einsteins vastly more valuable than they have ever been, even in percentage terms. But when I've looked at AI research I just haven't seen it. I've seen innumerable permutations on a few recycled architectural concepts, and way too much energy wasted on architectures in systems starved for content, content that

academic researchers have little incentive to pursue. So we have come to: What evidence is there for a dense sequence of powerful architectural AI insights? Is there any evidence that natural selection stumbled across such things?

And if Eliezer is the outlier he seems on the priority of friendly AI, what does Eliezer know that the rest of us don't? If he has such revolutionary clues, why can't he tell us? What else could explain his confidence and passion here if not such clues?

* * *

Eliezer Yudkowsky

> On small scales we humans evolved to cooperate via various pair and group bonding mechanisms. But these mechanisms aren't of much use on today's evolutionarily unprecedented large scales. Yet we do in fact cooperate on the largest scales. We do this because we are risk averse, because our values mainly conflict on resource use which conflicts destroy, and because we have the intelligence and institutions to enforce win-win deals via property rights, etc.

Individual organisms are adaptation-executers, not fitness-maximizers. We seem to have a disagreement-of-fact here; I think that our senses of honor and of internalized group morality are operating to make us honor our agreements with trade partners and internalize certain capitalist values. If human beings were *really genuinely* selfish, the economy would fall apart or at least have to spend vastly greater resources policing itself—think Zimbabwe and other failed states where police routinely stop buses to collect bribes from all passengers, but without the sense of restraint: the police just shoot you and loot your corpse unless they expect to be able to extract further bribes from you in particular.

I think the group coordination mechanisms, executing as adaptations, are *critical* to the survival of a global economy between imperfect minds of our

level that cannot simultaneously pay attention to everyone who might betray us.

> In this case the whole AI sector of the economy might grow very quickly, but it gets pretty hard to imagine one AI project zooming vastly ahead of others.

Robin, you would seem to be leaving out a key weak point here. It's much easier to argue that AIs don't zoom ahead of each other than to argue that the AIs as a *collective* don't zoom ahead of the *humans*. To the extent where, if AIs lack innate drives to treasure sentient life and humane values, it would be a trivial coordination problem and a huge net benefit to all AIs to simply write the statue-slow, defenseless, noncontributing humans out of the system.

Robin Hanson

> Eliezer: If human beings were *really genuinely* selfish, the economy would fall apart or at least have to spend vastly greater resources policing itself. . . . Group coordination mechanisms, executing as adaptations, are *critical* to the survival of a global economy. . . . It would be a trivial coordination problem and a huge net benefit to all AIs to simply write the statue-slow, defenseless, noncontributing humans out of the system.

Here you disagree with most economists, including myself, about the sources and solutions of coordination problems. Yes, genuinely selfish humans would have to spend more resources to coordinate at the local level, because this is where adapted coordinations now help. But larger-scale coordination would be just as easy. Since coordination depends crucially on institutions, AIs would need to preserve those institutions as well. So AIs would not want to threaten the institutions they use to keep the peace among themselves. It is far from easy to coordinate to exterminate humans while preserving such institutions. Also, why assume AIs not explicitly designed to be friendly are in fact "really genuinely selfish"?

See original post for all comments.

47

True Sources of Disagreement

Eliezer Yudkowsky

8 December 2008

Followup to: Is That Your True Rejection?

I expected from the beginning that the difficult part of two rationalists reconciling a persistent disagreement, would be for them to expose the true sources of their beliefs.

One suspects that this will only work if each party takes responsibility for their own end; it's very hard to see inside someone else's head. Yesterday I exhausted myself mentally while out on my daily walk, asking myself the Question "What do you think you know, and why do you think you know it?" with respect to "How much of the AI problem compresses to large insights, and how much of it is unavoidable nitty-gritty?" Trying to either understand why my brain believed what it believed, or else force my brain to experience enough genuine

doubt that I could reconsider the question and arrive at a real justification that way. It's hard to see how Robin Hanson could have done any of this work for me.

Presumably a symmetrical fact holds about my lack of access to the real reasons why Robin believes what he believes. To understand the true source of a disagreement, you have to know why *both* sides believe what they believe—one reason why disagreements are hard to resolve.

Nonetheless, here's my guess as to what this Disagreement is about:

If I had to pinpoint a single thing that strikes me as "disagreeable" about the way Robin frames his analyses, it's that there are a lot of *opaque* agents running around, little black boxes assumed to be similar to humans, but there are more of them and they're less expensive to build/teach/run. They aren't even any *faster*, let alone smarter. (I don't think that standard economics says that doubling the population halves the doubling time, so it matters whether you're making more minds or faster ones.)

This is Robin's model for uploads/ems, and his model for AIs doesn't seem to look any different. So that world looks like this one, except that the cost of "human capital" and labor is dropping according to (exogenous) Moore's Law, and it ends up that economic growth doubles every month instead of every sixteen years—but that's it. Being, myself, not an economist, this *does* look to me like a viewpoint with a distinctly economic zeitgeist.

In my world, you look inside the black box. (And, to be symmetrical, I don't spend much time thinking about more than one box at

a time—if I have more hardware, it means I have to figure out how to scale a bigger brain.)

The human brain is a haphazard thing, thrown together by idiot evolution as an incremental layer of icing on a chimpanzee cake that never evolved to be generally intelligent, adapted in a distant world devoid of elaborate scientific arguments or computer programs or professional specializations.

It's amazing we can get *anywhere* using the damn thing. But it's worth remembering that if there were any *smaller* modification of a chimpanzee that spontaneously gave rise to a technological civilization, we would be having this conversation at that lower level of intelligence instead.

Human neurons run at less than a millionth the speed of transistors, transmit spikes at less than a millionth the speed of light, and dissipate around a million times the heat per synaptic operation as the thermodynamic minimum for a one-bit operation at room temperature. Physically speaking, it ought to be possible to run a brain at a million times the speed without shrinking it, cooling it, or invoking reversible computing or quantum computing.

There's no reason to think that the brain's software is any closer to the limits of the possible than its hardware, and indeed, if you've been following along on *Overcoming Bias* this whole time, you should be well aware of the manifold known ways in which our high-level thought processes fumble even the simplest problems.

Most of these are not deep, inherent flaws of intelligence, or limits of what you can do with a mere hundred trillion computing elements. They are the results of a really stupid process that designed the retina

backward, slapping together a brain we now use in contexts way outside its ancestral environment.

Ten thousand researchers working for one year cannot do the same work as a hundred researchers working for a hundred years; a chimpanzee's brain is one-fourth the volume of a human's but four chimps do not equal one human; a chimpanzee shares 95% of our DNA but a chimpanzee cannot understand 95% of what a human can. The scaling law for population is not the scaling law for time is not the scaling law for brain size is not the scaling law for mind design.

There's a parable I sometimes use, about how the first replicator was not quite the end of the era of stable accidents, because the pattern of the first replicator was, of necessity, something that could happen by accident. It is only the *second* replicating pattern that you would never have seen without many copies of the first replicator around to give birth to it; only the *second* replicator that was part of the world of evolution, something you wouldn't see in a world of accidents.

That first replicator must have looked like one of the most bizarre things in the whole history of time—this *replicator* created purely by *chance*. But the history of time could never have been set in motion, otherwise.

And what a bizarre thing a human must be, a mind born entirely of evolution, a mind that was not created by another mind.

We haven't yet *begun* to see the shape of the era of intelligence.

Most of the universe is far more extreme than this gentle place, Earth's cradle. Cold vacuum or the interior of stars—either is far more common than the temperate weather of Earth's surface, where life first arose, in the balance between the extremes. And most possible intelli-

gences are not balanced, like these first humans, in that strange small region of temperate weather between an amoeba and a Jupiter Brain.

This is the challenge of my own profession—to break yourself loose of the tiny human dot in mind-design space, in which we have lived our whole lives, our imaginations lulled to sleep by too-narrow experiences.

For example, Robin says:

> *Eliezer guesses that within a few weeks a single AI could grow via largely internal means from weak and unnoticed to so strong it takes over the world.* [his italics]

I suppose that to a human a "week" sounds like a temporal constant describing a "short period of time," but it's actually 10^{49} Planck intervals, or enough time for a population of 2 GHz processor cores to perform 10^{15} *serial* operations one after the other.

Perhaps the thesis would sound less shocking if Robin had said, "Eliezer guesses that 10^{15} sequential operations might be enough to . . ."

One should also bear in mind that the human brain, which is not designed for the primary purpose of scientific insights, does not spend its power efficiently on having many insights in minimum time, but this issue is harder to understand than CPU clock speeds.

Robin says he doesn't like "unvetted abstractions." Okay. That's a strong point. I get it. Unvetted abstractions go kerplooie, yes they do indeed. But something's wrong with using that as a justification for models where there are lots of little black boxes just like humans scurrying around and we never pry open the black box and scale the brain bigger or redesign its software or even just *speed up* the damn thing.

The interesting part of the problem is *harder to analyze*, yes—more distant from the safety rails of overwhelming evidence—but this is no excuse for *refusing to take it into account*.

And in truth I do suspect that a strict policy against "unvetted abstractions" is not the real issue here. I constructed a simple model of an upload civilization running on the computers their economy creates: If a nonupload civilization has an exponential Moore's Law, $y = e^t$, then, naively, an upload civilization ought to have $\frac{dy}{dt} = e^y \rightarrow y = -\ln(C - t)$. *Not* necessarily up to infinity, but for as long as Moore's Law would otherwise stay exponential in a biological civilization. I walked though the implications of this model, showing that in many senses it behaves "just like we would expect" for describing a civilization running on its own computers.

Compare this to Robin Hanson's "Economic Growth Given Machine Intelligence",[1] which Robin describes as using "one of the simplest endogenous growth models to explore how Moore's Law changes with computer-based workers. It is an early but crude attempt, but it is the sort of approach I think promising." Take a quick look at that paper.

Now, consider the *abstractions* used in my Moore's Researchers scenario, versus the *abstractions* used in Hanson's paper above, and ask yourself *only* the question of which looks more "vetted by experience"—given that both are models of a sort that haven't been used before, in domains not actually observed, and that both give results quite different from the world we see—and that would probably cause the vast majority of actual economists to say, "Naaaah."

Moore's Researchers versus "Economic Growth Given Machine Intelligence"—if you didn't think about the *conclusions* in advance of

the reasoning; and if you also neglected that one of these has been written up in a way that is more impressive to economics journals; and you just asked the question, "To what extent is the math used here, constrained by our prior experience?" then I would think that the race would at best be even. Or possibly favoring "Moore's Researchers" as being more simple and intuitive, and involving less novel math as measured in additional quantities and laws introduced.

I ask in all humility if Robin's true rejection is a strictly evenhandedly applied rule that rejects unvetted abstractions. Or if, in fact, Robin finds my conclusions, and the sort of premises I use, to be *objectionable for other reasons*—which, so far as we know at this point, may well be *valid* objections—and so it appears to him that my abstractions bear *a larger burden of proof* than the sort of mathematical steps he takes in "Economic Growth Given Machine Intelligence." But rather than offering the reasons why the burden of proof appears larger to him, he says instead that it is "not vetted enough."

One should understand that "Your abstractions are unvetted!" makes it difficult for me to engage properly. The core of my argument has to do with what happens when you pry open the black boxes that are your economic agents, and start fiddling with their brain designs, and leave the tiny human dot in mind-design space. If all such possibilities are rejected *on the basis of their being "unvetted" by experience*, it doesn't leave me with much to talk about.

Why not just accept the rejection? Because I expect that to give the wrong answer—I expect it to ignore the dominating factor in the Future, even if the dominating factor is harder to analyze.

It shouldn't be surprising if a persistent disagreement ends up resting on that point where your attempt to take into account the

other person's view runs up against some question of simple fact where, it *seems* to you, *you know that can't possibly be right*.

For me, that point is reached when trying to visualize a model of interacting black boxes that behave like humans except they're cheaper to make. The world, which shattered once with the first replicator, and shattered for the second time with the emergence of human intelligence, somehow does *not* shatter a third time. Even in the face of blowups of brain size far greater than the size transition from chimpanzee brain to human brain; and changes in design far larger than the design transition from chimpanzee brains to human brains; and simple serial thinking speeds that are, maybe even right from the beginning, thousands or millions of times faster.

That's the point where I, having spent my career trying to look inside the black box, trying to wrap my tiny brain around the rest of mind-design space that isn't like our small region of temperate weather, just can't make myself believe that the Robin-world is *really truly actually* the way the future will be.

There are other things that seem like probable nodes of disagreement:

Robin Hanson's description of Friendly AI development as "total war" that is harmful to even discuss, or his description of a realized Friendly AI as "a God to rule us all." Robin must be visualizing an in-practice outcome very different from what I do, and this seems like a likely source of emotional fuel for the disagreement as well.

Conversely, Robin Hanson seems to approve of a scenario where lots of AIs, of arbitrary motives, constitute the vast part of the economic productivity of the Solar System, because he thinks that humans will be protected under the legacy legal system that grew con-

tinuously out of the modern world, and that the AIs will be unable to coordinate to transgress the legacy legal system for fear of losing their own legal protections. I tend to visualize a somewhat different outcome, to put it mildly, and would symmetrically be suspected of emotional unwillingness to accept that outcome as inexorable.

Robin doesn't dismiss Cyc out of hand and even "hearts" it, which implies that we have extremely different pictures of how intelligence works.

Like Robin, I'm also feeling burned on this conversation, and I doubt we'll finish it; but I should write at least two more posts to try to describe what I've learned, and some of the rules that I think I've been following.

* * *

Robin Hanson

Miscellaneous points:

- I guessed a week to month doubling time, not six months.
- I've talked explicitly about integrated communities of faster ems.
- I used a learning-by-doing modeling approach to endogenize Moore's Law.
- Any model of minds usable for forecasting world trends must leave out detail.
- Most people complain that economists using game theory to model humans ignore too much human detail; what *excess* human detail do you think economists retain?

- Research labs hiring workers, e.g., Intel, are willing to trade off worker speed, i.e., hours per week, for worker salary, experience, etc.; a model that says Intel cares only about worker speed misses an awful lot.

Eliezer Yudkowsky

Robin, I found different guesses at the doubling time listed in different places, so I just used one from "Economic Growth Given Machine Intelligence." I'll change the text.

Robin Hanson

. . . Eliezer, most readers of this blog are not in a position to evaluate which model looks more vetted. The whole point is that a community of thousands of specialists has developed over decades vetting models of total system growth, and they are in the best position to judge. I have in fact not just talked about vetting, but have offered more detailed reasons why your model seems unsatisfactory.

Eliezer Yudkowsky

. . . Robin, should we ask James Miller then? I have no problem with the detailed reasons you offer, it's just the "insufficiently vetted" part of the argument that I find difficult to engage with—unless I actually find members of this community and ask them which specific pieces are "vetted" in their view, by what evidence, and which not. I wouldn't necessarily trust them, to be frank, because it was never a condition of their profession that they should deal with nonhumans. But at least I would have some idea of what those laws were under which I was being judged.

It's hard for me to accept as normative the part of this argument that is an appeal to authority (professional community that has learned good norms about constructing growth models) rather than an appeal to evidence (look at how well the evidence fits these specific growth models). It's not that I reject authority in general, but these people's professional experience is entirely about humans, and it's hard for me to believe that they have taken into account the considerations involved in extrapolating narrow experience to non-narrow experience when various basic assumptions are potentially broken. I would expect them to have norms that worked for describing humans, full stop.

Robin Hanson

Eliezer, I'm not sure James Miller has done much econ growth research. How about my colleague Garrett Jones, who specializes in intelligence and growth?

Eliezer Yudkowsky

Robin, I'd be interested, but I'd ask whether you've discussed this particular issue with Jones before. (I.e., the same reason I don't cite Peter Cheeseman as support for, e.g., the idea that *general* AI mostly doesn't work if you don't have all the parts, and then undergoes something like a chimp → human transition as soon as all the parts are in place. So far as I can tell, Cheeseman had this idea before I met him; but he still wouldn't be an unbiased choice of referee, because I already know many of his opinions and have explicitly contaminated him on some points.)

Robin Hanson

Eliezer, Garrett has seen and likes my growth paper, but he and I have not talked at all about your concepts. I sent him a link once to this post of yours;[2] I'll email you his reply.

Eliezer Yudkowsky

. . . Robin, email reply looks fine.

See original post for all comments.

* * *

1. Hanson, "Economic Growth Given Machine Intelligence."

2. Eliezer Yudkowsky, "Economic Definition of Intelligence?," *Less Wrong* (blog), October 29, 2008, http://lesswrong.com/lw/vc/economic_definition_of_intelligence/.

48

The Bad Guy Bias

Robin Hanson

9 December 2008

Shankar Vedantam:

> Nations tend to focus far more time, money and attention on tragedies caused by human actions than on the tragedies that cause the greatest amount of human suffering or take the greatest toll in terms of lives. . . . In recent years, a large number of psychological experiments have found that when confronted by tragedy, people fall back on certain mental rules of thumb, or heuristics, to guide their moral reasoning. When a tragedy occurs, we instantly ask who or what caused it. When we find a human hand behind the tragedy—such as terrorists, in the case of the Mumbai attacks—something clicks in our minds that makes the tragedy seem worse than if it had been caused by an act of nature, disease or even human apathy. . . .

> Tragedies, in other words, cause individuals and nations to behave a little like the detectives who populate television murder mystery shows: We spend nearly all our time on the victims of killers and rapists and very little on the victims of car accidents and smoking-related lung cancer.
>
> "We think harms of actions are much worse than harms of omission," said Jonathan Baron, a psychologist at the University of Pennsylvania. "We want to punish those who act and cause harm much more than those who do nothing and cause harm. We have more sympathy for the victims of acts rather than the victims of omission. If you ask how much should victims be compensated, [we feel] victims harmed through actions deserve higher compensation."[1]

This bias should also afflict our future thinking, making us worry more about evil alien intent than unintentional catastrophe.

* * *

Eliezer Yudkowsky

Indeed, I've found that people repeatedly ask me about AI projects with ill intentions—Islamic terrorists building an AI—rather than trying to grasp the ways that well-intentioned AI projects go wrong by default.

See original post for all comments.

* * *

1. Shankar Vedantam, "In Face of Tragedy, 'Whodunit' Question Often Guides Moral Reasoning," *Washington Post*, December 8, 2008, accessed November 25, 2012, http:

//www.washingtonpost.com/wp-dyn/content/article/2008/12/07/AR2008120702830.html.

49

Disjunctions, Antipredictions, Etc.

Eliezer Yudkowsky

9 December 2008

Followup to: Underconstrained Abstractions

Previously:

> So if it's not as simple as *just* using the one trick of finding abstractions you can easily verify on available data . . . what are some other tricks to use?

There are several, as you might expect . . .

Previously I talked about "permitted possibilities." There's a trick in debiasing that has mixed benefits, which is to try and visualize several specific possibilities instead of just one.

The reason it has "mixed benefits" is that being specific, at all, can have biasing effects relative to just imagining a typical case. (And be-

lieve me, if I'd seen the outcome of a hundred planets in roughly our situation, I'd be talking about that instead of all this Weak Inside View stuff.)

But if you're going to bother visualizing the future, it does seem to help to visualize more than one way it could go, instead of concentrating all your strength into *one* prediction.

So I try not to ask myself, "What will happen?" but rather, "Is this possibility allowed to happen, or is it prohibited?" There are propositions that seem forced to me, but those should be relatively rare—the first thing to understand about the future is that it is hard to predict, and you shouldn't seem to be getting strong information about most aspects of it.

Of course, if you allow more than one possibility, then you have to discuss more than one possibility, and the total length of your post gets longer. If you just eyeball the length of the post, it looks like an unsimple theory; and then talking about multiple possibilities makes you sound weak and uncertain.

As Robyn Dawes notes,

> In their summations lawyers avoid arguing from disjunctions in favor of conjunctions. (There are not many closing arguments that end, "Either the defendant was in severe financial straits and murdered the decedent to prevent his embezzlement from being exposed or he was passionately in love with the same coworker and murdered the decedent in a fit of jealous rage or the decedent had blocked the defendant's promotion at work and the murder was an act of revenge. The State has given you solid evidence to support each of these alternatives, all of which would lead to the same conclusion: first-degree murder.") Rationally,

> of course, disjunctions are much *more* probable than are conjunctions.[1]

Another test I use is simplifiability—*after* I've analyzed out the idea, can I compress it *back* into an argument that fits on a T-shirt, even if it loses something thereby? Here's an example of some compressions:

- The whole notion of recursion and feeding object-level improvements back into meta-level improvements: "If computing power per dollar doubles every eighteen months, what happens if computers are doing the research?"
- No diminishing returns on complexity in the region of the transition to human intelligence: "We're so similar to chimps in brain design, and yet so much more powerful; the upward slope must be really steep."
- Scalability of hardware: "Humans have only four times the brain volume of chimps—now imagine an AI suddenly acquiring a thousand times as much power."

If the whole argument was that T-shirt slogan, I wouldn't find it compelling—too simple and surface a metaphor. So you have to look more closely, and try visualizing some details, and make sure the argument can be consistently realized so far as you know. But if, *after* you do that, you can compress the argument back to fit on a T-shirt again—even if it sounds naive and stupid in that form—then that helps show that the argument doesn't *depend* on all the details being true simultaneously; the details might be different while fleshing out the same core idea.

Note also that the three statements above are to some extent disjunctive—you can imagine only one of them being true, but a hard takeoff still occurring for just that reason alone.

Another trick I use is the idea of *antiprediction*. This is when the narrowness of our human experience distorts our metric on the answer space, and so you can make predictions that actually aren't far from max-entropy priors, but *sound* very startling.

I shall explain:

A news story about an Australian national lottery that was just starting up, interviewed a man on the street, asking him if he would play. He said yes. Then they asked him what he thought his odds were of winning. "Fifty–fifty," he said, "either I win or I don't."

To predict your odds of winning the lottery, you should invoke the Principle of Indifference with respect to all possible combinations of lottery balls. But this man was invoking the Principle of Indifference with respect to the partition "win" and "not win." To him, they sounded like equally simple descriptions; but the former partition contains only one combination, and the latter contains the other N million combinations. (If you don't agree with this analysis, I'd like to sell you some lottery tickets.)

So the *antiprediction* is just "You won't win the lottery." And the one may say, "What? How do you know that? You have no evidence for that! You can't prove that I won't win!" So they are focusing far too much attention on a small volume of the answer space, artificially inflated by the way their attention dwells upon it.

In the same sense, if you look at a television SF show, you see that a remarkable number of aliens seem to have human body plans—two arms, two legs, walking upright, right down to five fingers per hand

and the location of eyes in the face. But this is a very narrow partition in the body-plan space; and if you just said, "They won't look like humans," that would be an antiprediction that just steps outside this artificially inflated tiny volume in the answer space.

Similarly with the true sin of television SF, which is too-human minds, even among aliens not meant to be sympathetic characters. "If we meet aliens, they won't have a sense of humor," I antipredict; and to a human it sounds like I'm saying something highly specific, because all minds by default have a sense of humor, and I'm predicting the presence of a no-humor attribute tagged on. But actually, I'm just predicting that a point in mind-design volume is outside the narrow hyperplane that contains humor.

An AI might go from infrahuman to transhuman in *less than a week*? But a week is 10^{49} Planck intervals—if you just look at the exponential scale that stretches from the Planck time to the age of the universe, there's nothing special about the timescale that 200 Hz humans happen to live on, any more than there's something special about the numbers on the lottery ticket you bought.

If we're talking about a starting population of 2 GHz processor cores, then any given AI that FOOMs at all is likely to FOOM in less than 10^{15} sequential operations or more than 10^{19} sequential operations, because the region between 10^{15} and 10^{19} isn't all that wide a target. So less than a week or more than a century, and in the latter case that AI will be trumped by one of a shorter timescale.

This is actually a pretty naive version of the timescale story. But as an example, it shows how a "prediction" that's close to just stating a maximum-entropy prior can sound amazing, startling, counterintuitive, and futuristic.

When I make an antiprediction supported by disjunctive arguments that are individually simplifiable, I feel *slightly* less nervous about departing the rails of vetted abstractions. (In particular, I regard this as sufficient reason not to trust the results of generalizations over only human experiences.)

Finally, there are three tests I apply to figure out how strong my predictions are.

The first test is to just ask myself the Question "What do you think you know, and why do you think you know it?" The future is something I haven't yet observed; if my brain claims to know something about it with any degree of confidence, what are the reasons for that? The first test tries to align the strength of my predictions with things that I have reasons to believe—a basic step, but one which brains are surprisingly wont to skip.

The second test is to ask myself, "How worried do I feel that I'll have to write an excuse explaining why this happened anyway?" If I don't feel worried about having to write an excuse—if I can stick my neck out and not feel too concerned about ending up with egg on my face—then clearly my brain really does believe this thing quite strongly, not as a point to be professed through enthusiastic argument, but as an ordinary sort of fact. Why?

And the third test is the "So what?" test—to what degree will I feel indignant if Nature comes back and says, "So what?" to my clever analysis? Would I feel as indignant as if I woke up one morning to read in the newspaper that Mars had started orbiting the Sun in squares instead of ellipses? Or, to make it somewhat less strong, as if I woke up one morning to find that banks were charging negative interest on loans? If so, clearly I must possess some kind of *ex-*

tremely strong argument—one that even Nature Itself ought to find compelling, not just humans. What is it?

* * *

See original post for all comments.

* * *

1. Robyn M. Dawes, *Rational Choice in An Uncertain World*, 1st ed., ed. Jerome Kagan (San Diego, CA: Harcourt Brace Jovanovich, 1988).

50

Are AIs Homo Economicus*?*

Robin Hanson

9 December 2008

Eliezer yesterday:

> If I had to pinpoint a single thing that strikes me as "disagree-able" about the way Robin frames his analyses, it's that there are a lot of *opaque* agents running around, little black boxes assumed to be similar to humans, but there are more of them and they're less expensive to build/teach/run. . . . The core of my argument has to do with what happens when you pry open the black boxes that are your economic agents, and start fiddling with their brain designs, and leave the tiny human dot in mind-design space.

Lots of folks complain about economists; believers in peak oil, the gold standard, recycling, electric cars, rent control, minimum wages, tariffs, and bans on all sorts of things complain about contrary economic analyses. Since compared to most social scientists economists

use relatively stark mathy models, the usual complaint is that our models neglect relevant factors and make false assumptions.

But of course we must neglect most everything, and make false assumptions, to have tractable models; the question in each context is what neglected factors and false assumptions would most mislead us.

It is odd to hear complaints that economic models assume too much humanity; the usual complaint is the opposite. Unless physicists have reasons to assume otherwise, they usually assume masses are at points, structures are rigid, surfaces are frictionless, and densities are uniform. Similarly, unless economists have reasons to be more realistic in a context, they usually assume people are identical, risk neutral, live forever, have selfish material stable desires, know everything, make no mental mistakes, and perfectly enforce every deal. Products usually last one period or forever, are identical or infinitely varied, etc.

Of course we often do have reasons to be more realistic, considering deals that may not be enforced; people who die; people with diverse desires, info, abilities, and endowments; people who are risk averse, altruistic, or spiteful; people who make mental mistakes; and people who follow "behavioral" strategies. But the point isn't just to add as much realism as possible; it is to be clever about knowing which sorts of detail are most relevant in what context.

So to a first approximation, economists can't usually tell if the agents in their models are AIs or human! But we can still wonder: how could economic models better capture AIs? In common with ems, AIs could make copies of themselves, save backups, and run at varied speeds. Beyond ems, AIs might buy or sell mind parts, and

reveal mind internals, to show commitment to actions or honesty of stated beliefs. Of course,

> That might just push our self-deception back to the process that produced those current beliefs. To deal with self-deception in belief production, we might want to provide audit trails, giving more transparency about the origins of our beliefs.[1]

Since economists feel they understand the broad outlines of cooperation and conflict pretty well using simple stark models, I am puzzled to hear Eliezer say:

> If human beings were *really genuinely* selfish, the economy would fall apart or at least have to spend vastly greater resources policing itself. . . . Group coordination mechanisms, executing as adaptations, are critical to the survival of a global economy.

We think we understand just fine how genuinely selfish creatures can cooperate. Sure, they might have to spend somewhat greater on policing, but not *vastly* greater, and a global economy could survive just fine. This seems an important point, as it seems to be why Eliezer fears even nonlocal AI fooms.

* * *

Eliezer Yudkowsky

The main part you're leaving out of your models (on my view) is the part where AIs can scale on hardware by expanding their brains, and scale on software by redesigning themselves, and these scaling curves are much sharper

than "faster" let alone "more populous." Aside from that, of course, AIs are more like economic agents than humans are.

My statement about "truly selfish humans" isn't meant to be about truly selfish AIs, but rather, truly selfish entities with limited human attention spans, who have much worse agent problems than an AI that can monitor all its investments simultaneously and inspect the source code of its advisers. The reason I fear nonlocal AI fooms is precisely that they would have no trouble coordinating to cut the legacy humans out of their legal systems.

Robin Hanson

Eliezer, economists assume that every kind of product can be improved, in terms of cost and performance, and we have many detailed models of product innovation and improvement. The hardware expansion and software redesign that you say I leave out seem to me included in the mind parts that can be bought or sold. How easy it is to improve such parts, and how much better parts add to mind productivity, is exactly the debate we've been having.

See original post for all comments.

* * *

1. Robin Hanson, "Enhancing Our Truth Orientation," in *Human Enhancement*, 1st ed., ed. Julian Savulescu and Nick Bostrom (New York: Oxford University Press, 2009), 257–274.

51

Two Visions Of Heritage

Robin Hanson

9 December 2008

Eliezer and I seem to disagree on our heritage.

I see our main heritage from the past as all the innovations embodied in the design of biological cells/bodies, of human minds, and of the processes/habits of our hunting, farming, and industrial economies. These innovations are mostly steadily accumulating modular "content" within our architectures, produced via competitive processes and implicitly containing both beliefs and values. Architectures also change at times as well.

Since older heritage levels grow more slowly, we switch when possible to rely on newer heritage levels. For example, we once replaced hunting processes with farming processes, and within the next century we may switch from bio to industrial mental hardware, becoming

ems. We would then rely far less on bio and hunting/farm heritages, though still lots on mind and industry heritages. Later we could make AIs by transferring mind content to new mind architectures. As our heritages continued to accumulate, our beliefs and values should continue to change.

I see the heritage we will pass to the future as mostly avoiding disasters to preserve and add to these accumulated contents. We might get lucky and pass on an architectural change or two as well. As ems we can avoid our bio death heritage, allowing some of us to continue on as ancients living on the margins of far future worlds, personally becoming a heritage to the future.

Even today one could imagine overbearing systems of property rights giving almost all income to a few. For example, a few consortiums might own every word or concept and require payments for each use. But we do not have such systems, in part because they would not be enforced. One could similarly imagine future systems granting most future income to a few ancients, but those systems would also not be enforced. Limited property rights, however, such as to land or sunlight, would probably be enforced just to keep peace among future folks, and this would give even unproductive ancients a tiny fraction of future income, plenty for survival among such vast wealth.

In contrast, it seems **Eliezer sees** a universe where In the Beginning arose a blind and indifferent but prolific creator, who eventually made a race of seeing creators, creators who could also love, and love well. His story of the universe centers on the loves and sights of a team of geniuses of mind design, a team probably alive today. This genius team will see deep into the mysteries of mind, far

deeper than all before, and learn to create a seed AI mind architecture which will suddenly, and with little warning or outside help, grow to take over the world. If they are wise, this team will also see deep into the mysteries of love, to make an AI that forever loves what that genius team wants it to love.

As the AI creates itself it reinvents everything from scratch using only its architecture and raw data; it has little need for other bio, mind, or cultural content. All previous heritage aside from the genius team's architecture and loves can be erased more thoroughly than the Biblical flood supposedly remade the world. And forevermore from that point on, the heritage of the universe would be a powerful unrivaled AI singleton, i.e., a God to rule us all, that does and makes what it loves.

If God's creators were wise then God is unwavering in loving what it was told to love; if they were unwise, then the universe becomes a vast random horror too strange and terrible to imagine. Of course other heritages may be preserved if God's creators told him to love them; and his creators would probably tell God to love themselves, their descendants, their associates, and their values.

The contrast between these two views of our heritage seems hard to overstate. One is a dry account of small individuals whose abilities, beliefs, and values are set by a vast historical machine of impersonal competitive forces, while the other is a grand inspiring saga of absolute good or evil hanging on the wisdom of a few mythic heroes who use their raw genius and either love or indifference to make a God who makes a universe in the image of their feelings. How does one begin to compare such starkly different visions?

* * *

Eliezer Yudkowsky

Needless to say, I don't think this represents my views even poorly, but to focus on your own summary:

> As our heritages continued to accumulate, our beliefs and values should continue to change.

You don't seem very upset about this "values change" process. Can you give an example of a values change that might occur? Are there values changes that you wouldn't accept, or that you would regard as an overwhelming disaster?

Naively, one would expect that a future in which very few agents share your utility function is a universe that will have very little utility from your perspective. Since you don't seem to feel that this is the case, are there things you value that you expect to be realized by essentially arbitrary future agents? What are these things?

What is it that your Future contains which is good, which you expect to be realized even if almost no one values this good in itself?

If the answer is "nothing" then the vision that you have sketched is of a universe empty of value; we should be willing to take almost any risk to prevent its realization.

> Even today one could imagine overbearing systems of property rights giving almost all income to a few. For example, a few consortiums might own every word or concept and require payments for each use. But we do not have such systems, in part because they would not be enforced. One could similarly imagine future systems granting most future income to a few ancients, but those systems would also not be enforced.

Please walk us through the process by which you think, if most future capital or income were granted to a few ancients under a legacy legal system, a poor majority of AIs would reject this legal system and replace it with something else. What exactly goes through their minds? How is the process of replacing the legacy legal system carried out?

Robin Hanson

. . . Eliezer, I'll correct errors you point out in views I attribute to you. This post is taking seriously your suggestion to look deeper for the core of our disagreement. My vision isn't of a universe as I want it to be, but of a universe as it is. An example of a future values change would be ems only mildly upset at death, when many other recent copies still live. I can see why they would have such values, and it doesn't seem a terrible thing to me. I'll consider writing a new post about rebellion against legacies.

See original post for all comments.

52

The Mechanics of Disagreement

Eliezer Yudkowsky

10 December 2008

Two ideal Bayesians cannot have common knowledge of disagreement; this is a theorem. If two rationalist wannabes have common knowledge of a disagreement between them, what could be going wrong?

The obvious interpretation of these theorems is that if you know that a cognitive machine is a rational processor of evidence, its beliefs become evidence themselves.

If you design an AI and the AI says, "This fair coin came up heads with 80% probability," then you know that the AI has accumulated evidence with an likelihood ratio of 4:1 favoring heads—because the AI only emits that statement under those circumstances.

It's not a matter of charity; it's just that this is how you think the other cognitive machine works.

And if you tell an ideal rationalist, "I think this fair coin came up heads with 80% probability," and they reply, "I now think this fair coin came up heads with 25% probability," and your sources of evidence are independent of each other, then you should accept this verdict, reasoning that (before you spoke) the other mind must have encountered evidence with a likelihood of 1:12 favoring tails.

But this *assumes* that the other mind also thinks that *you're* processing evidence correctly, so that, by the time it says "I now think this fair coin came up heads, $p = .25$," it has already taken into account the full impact of all the evidence you know about, before adding more evidence of its own.

If, on the other hand, the other mind doesn't trust your rationality, then it won't accept your evidence at face value, and the estimate that it gives won't integrate the full impact of the evidence you observed.

So does this mean that when two rationalists trust each other's rationality less than completely, then they can agree to disagree?

It's not that simple. Rationalists should not trust *themselves* entirely, either.

So when the other mind accepts your evidence at less than face value, this doesn't say, "You are less than a perfect rationalist," it says, "I trust you less than you trust yourself; I think that you are discounting your own evidence too little."

Maybe your raw arguments seemed to you to have a strength of 40:1, but you discounted for your own irrationality to a strength of 4:1, but the other mind thinks you still overestimate yourself and so it assumes that the actual force of the argument was 2:1.

And if you *believe* that the other mind is discounting you in this way, and is unjustified in doing so, then when it says, "I now think

this fair coin came up heads with 25% probability," you might bet on the coin at odds of 57% in favor of heads—adding up your further-discounted evidence of 2:1 to the implied evidence of 1:6 that the other mind must have seen to give final odds of 2:6—*if* you even fully trust the other mind's further evidence of 1:6.

I think we have to be very careful to avoid interpreting this situation in terms of anything like a *reciprocal trade*, like two sides making *equal concessions* in order to reach agreement on a business deal.

Shifting beliefs is not a concession that you make for the sake of others, expecting something in return; it is an advantage you take for your own benefit, to improve your own map of the world. I am, generally speaking, a Millie-style altruist; but when it comes to *belief shifts* I espouse a pure and principled selfishness: don't believe you're doing it for anyone's sake but your own.

Still, I once read that there's a principle among con artists that the main thing is to get the mark to believe that *you trust them*, so that they'll feel obligated to trust you in turn.

And—even if it's for completely different theoretical reasons—if you want to persuade a rationalist to shift belief to match yours, you either need to persuade them that you have all of the same evidence they do and have already taken it into account, or that you already fully trust their opinions as evidence, or that you know better than they do how much they themselves can be trusted.

It's that last one that's the really sticky point, for obvious reasons of asymmetry of introspective access and asymmetry of motives for overconfidence—how do you resolve that conflict? (And if you started *arguing* about it, then the question wouldn't be which of these were more important as a factor, but rather, which of these factors

the Other had under- or overdiscounted in forming their estimate of a given person's rationality . . .)

If I had to name a single reason why two wannabe rationalists wouldn't actually be able to agree in practice, it would be that once you trace the argument to the meta level where theoretically everything can be and must be resolved, the argument trails off into psychoanalysis and noise.

And if you look at what goes on in *practice* between two arguing rationalists, it would probably mostly be trading object-level arguments; and the most meta it would get is trying to convince the other person that you've already taken their object-level arguments into account.

Still, this does leave us with three clear reasons that someone might point to, to justify a persistent disagreement—even though the frame of mind of *justification* and having clear reasons to *point to* in front of others is itself antithetical to the spirit of resolving disagreements—but even so:

- *Clearly*, the Other's object-level arguments are flawed; no amount of trust that I can have for another person will make me believe that rocks fall upward.
- *Clearly*, the Other is not taking my arguments into account; there's an obvious asymmetry in how well I understand them and have integrated their evidence, versus how much they understand me and have integrated mine.

- *Clearly*, the Other is completely biased in how much they trust themselves over others, versus how I humbly and evenhandedly discount my own beliefs alongside theirs.

Since we don't want to go around encouraging disagreement, one might do well to ponder how all three of these arguments are used by creationists to justify their persistent disagreements with scientists.

That's one reason I say *clearly*—if it isn't obvious even to outside onlookers, maybe you shouldn't be confident of resolving the disagreement there. Failure at any of these levels implies failure at the meta-levels above it, but the higher-order failures might not be *clear*.

* * *

Robin Hanson

Of course if you knew that your disputant would only disagree with you when one of these three conditions clearly held, you would take their persistent disagreement as showing one of these conditions held, and then back off and stop disagreeing. So to apply these conditions you need the additional implicit condition that they do not believe that you could only disagree under one of these conditions.

See original post for all comments.

Part III

Conclusion

53

What Core Argument?

Robin Hanson

10 December 2008

People keep asking me to return to the core of the argument, but, well, there's just not much there. Let's review, again. Eliezer suggests someone soon may come up with a seed AI architecture allowing a single AI to within roughly a week grow from unimportant to strong enough to take over the world. I'd guess we are talking over twenty orders of magnitude growth in its capability, or sixty doublings.

This amazing growth rate sustained over such a large magnitude range is far beyond what the vast majority of AI researchers, growth economists, or most any other specialists would estimate. It is also far beyond estimates suggested by the usual choices of historical analogs or trends. Eliezer says the right reference set has two other elements, the origin of life and the origin of human minds, but why should we

accept this reference? He also has a math story to suggest this high average growth, but I've said:

> I also find Eliezer's growth math unpersuasive. Usually dozens of relevant factors are coevolving, with several loops of all else equal X growth speeds Y growth speeds etc. Yet usually it all adds up to exponential growth, with rare jumps to faster growth rates. Sure, if you pick two things that plausibly speed each other and leave everything else out including diminishing returns, your math can suggest accelerating growth to infinity, but for a real foom that loop needs to be real strong, much stronger than contrary muting effects.

Eliezer has some story about how chimp vs. human brain sizes shows that mind design doesn't suffer diminishing returns or low-hanging-fruit-first slowdowns, but I have yet to comprehend this argument. Eliezer says it is a myth that chip developers need the latest chips to improve chips as fast as they do, so there aren't really diminishing returns there, but chip expert Jed Harris seems to disagree.

Monday Eliezer said:

> Yesterday I exhausted myself . . . asking . . . "What do you think you know, and why do you think you know it?" with respect to, "How much of the AI problem compresses to large insights, and how much of it is unavoidable nitty-gritty?"

His answer:

> The human brain is a haphazard thing, thrown together by idiot evolution. . . . If there were any *smaller* modification of a chimpanzee that spontaneously gave rise to a technological civilization, we would be having this conversation at that lower level of intelligence instead.

> Human neurons run at less than a millionth the speed of transistors. . . . There's no reason to think that the brain's software is any closer to the limits of the possible than its hardware. . . . [Consider] the manifold known ways in which our high-level thought processes fumble even the simplest problems. Most of these are not deep, inherent flaws of intelligence. . . .
>
> We haven't yet *begun* to see the shape of the era of intelligence. Most of the universe is far more extreme than this gentle place, Earth's cradle. . . . Most possible intelligences are not balanced, like these first humans, in that strange small region of temperate weather between an amoeba and a Jupiter Brain. . . . I suppose that to a human a "week" sounds like a temporal constant describing a "short period of time," but it's actually 10^{49} Planck intervals.

I feel like the woman in Monty Python's "Can we have your liver?" sketch, cowed into giving her liver after hearing how vast is the universe. Sure, evolution being stupid suggests there are substantial architectural improvements to be found. *But that says nothing about the relative contribution of architecture and content in minds, nor does it say anything about how easy it will be to quickly find a larger number of powerful architectural improvements!*

* * *

Eliezer Yudkowsky

The question "How compressible is it?" is not related to the paragraph you quote. It is simply what I actually happened to be doing that day.

Twenty orders of magnitude in a week doesn't sound right, unless you're talking about the tail end *after* the AI gets nanotechnology. Figure more like some number of years to push the AI up to a critical point, two to six orders of magnitude improvement from there to nanotech, then some more orders of magnitude after that.

Eliezer Yudkowsky

Also, the notion is not that mind design never runs into diminishing returns. Just that you don't hit that point up to human intelligence. The main easily accessible arguments for why you don't hit diminishing returns for some time *after* human intelligence has to do with the idea that there's (a) nothing privileged about human intelligence and (b) lots of visible flaws in it.

Robin Hanson

I don't understand why visible flaws implies a lack of diminishing returns near the human level.

Eliezer Yudkowsky

It means you can go on past human *just* by correcting the flaws. If you look at the actual amount of cognitive work that we devote to the key insights in science, as opposed to chasing red herrings, clinging to silly ideas, or going to the bathroom, then there's at least three orders of magnitude speedup right there, I'd say, on the cognitive part of the process.

Robin Hanson

I'm talking orders of magnitude in total capacity to do things, something like economic product, because that seems the simplest overall metric. If the world has ten orders of magnitude of humans, then something that can take over the world is roughly that much bigger than a human. And presumably this AI starts as far less capable than a human. If this scenario happens in an em world, there'd be lots more stronger creatures to beat.

Eliezer, I don't see how that follows *at all*. Just because I can tell that a car's bumper is too heavy doesn't mean I have any idea how to make a car. You need to make a direct and clear argument. . . .

See original post for all comments.

54

What I Think, If Not Why

Eliezer Yudkowsky

11 December 2008

Reply to: Two Visions of Heritage

Though it really goes tremendously against my grain—it feels like sticking my neck out over a cliff (or something)—I guess I have no choice here but to try and make a list of *just* my positions, without justifying them. We can only talk justification, I guess, after we get straight what my positions *are*. I will also leave off many disclaimers to present the points *compactly* enough to be remembered.

- A well-designed mind should be ***much more efficient*** **than a human**, capable of doing more with less sensory data and fewer computing operations. It is not *infinitely efficient* and **does not use *zero* data**. But it does use little enough that *local pipelines*

such as a small pool of programmer-teachers, and later a huge pool of e-data, are sufficient.

- An AI that reaches a certain point in its own development becomes able to (sustainably, strongly) improve itself. At this point, recursive cascades **slam over many internal growth curves to near the limits of their current hardware**, and the AI undergoes a vast increase in capability. This point is at, or probably considerably before, a minimally transhuman mind capable of writing its own AI theory textbooks—an upper bound beyond which it could swallow and improve its *entire* design chain.

- It is *likely* that this capability increase or "FOOM" has an intrinsic maximum velocity that a human would regard as "fast" if it happens at all. A human week is $\sim 10^{15}$ serial operations for a population of 2 GHz cores, and a century is $\sim 10^{19}$ serial operations; this whole range is a narrow window. However, the core argument does not require one-week speed and a FOOM that takes two years ($\sim 10^{17}$ serial ops) will still carry the weight of the argument.

- **The *default* case of FOOM is an unFriendly AI, built by researchers with shallow insights**. This AI becomes able to improve itself in a haphazard way, makes various changes that are net improvements but may introduce value drift, and then gets smart enough to do guaranteed self-improvement, at which point its values freeze (forever).

- **The *desired* case of FOOM is a Friendly AI**, built using deep insight, so that the AI never makes any changes to itself that potentially change its internal values; all such changes are guaranteed using strong techniques that allow for a billion sequential self-modifications without losing the guarantee. The guarantee is written over the AI's *internal search criterion* for actions, rather than *external consequences*.

- The **good guys do *not* write** an AI which values **a bag of things that the programmers think are good ideas**, like libertarianism or socialism or making people happy or whatever. There were multiple *Less Wrong* sequences about this *one point*, like the Fake Utility Function sequence and the sequence on metaethics. It is dealt with at length in the document Coherent Extrapolated Volition. It is the first thing, the last thing, and the middle thing that I say about Friendly AI. I have said it over and over. I truly do not understand how anyone can pay *any* attention to *anything* I have said on this subject and come away with the impression that I think programmers are supposed to directly impress their nonmeta personal philosophies onto a Friendly AI.

- **The good guys do not directly impress their personal values onto a Friendly AI.**

- Actually setting up a Friendly AI's values is **an extremely *meta* operation,** less "make the AI want to make people happy" and more like "**superpose** the possible **reflective equilibria** of the **whole human species**, and **output new code** that over-

writes the current AI and has the **most coherent** support within that superposition."[1] This actually seems to be something of a *pons asinorum* in FAI—the ability to understand and endorse metaethical concepts that do not *directly* sound like amazing wonderful happy ideas. **Describing this as declaring total war on the rest of humanity does not seem fair** (or accurate).

- **I myself am strongly individualistic:** The most painful memories in my life have been when other people thought they knew better than me, and tried to do things on my behalf. It is also a known principle of hedonic psychology that people are happier when they're steering their own lives and doing their own interesting work. When I try myself to visualize what a beneficial superintelligence ought to do, it consists of **setting up a world that works by better rules, and then fading into the background,** silent as the laws of Nature once were, and finally folding up and vanishing when it is no longer needed. But this is only the thought of my mind that is merely human, and **I am barred from programming any such consideration *directly* into a Friendly AI,** for the reasons given above.

- Nonetheless, it does seem to me that this particular scenario **could not be justly described as "a God to rule over us all,"** unless the current fact that humans age and die is "a malevolent God to rule us all." So either Robin has a very different idea about what human reflective equilibrium values are likely to look like; or Robin believes that the Friendly AI project is bound to *fail* in such way as to create a paternalistic God; or—and this seems more likely to me—Robin didn't read all the way

through all the blog posts in which I tried to explain all the ways that this is not how Friendly AI works.

- **Friendly AI is technically difficult and requires an extraordinary effort on multiple levels.** English sentences like "make people happy" cannot describe the values of a Friendly AI. Testing is not sufficient to guarantee that values have been successfully transmitted.

- White-hat AI researchers are distinguished by the degree to which **they understand that a single misstep could be fatal, and can discriminate strong and weak assurances.** Good intentions are not only common, they're cheap. The story isn't about good versus evil, it's about people trying to do the impossible versus others who . . . aren't.

- Intelligence is about being able to **learn lots of things, not about knowing lots of things.** Intelligence is especially not about tape-recording lots of parsed English sentences à la Cyc. Old AI work was poorly focused due to inability to introspectively see the first and higher *derivatives* of knowledge; human beings have an easier time reciting sentences than reciting their ability to learn.

- **Intelligence is mostly about architecture,** or "knowledge" along the lines of knowing to look for causal structure (Bayes-net type stuff) in the environment; this kind of knowledge will usually be expressed procedurally as well as declaratively. **Architecture is mostly about deep insights.** This point has not yet been addressed (much) on *Overcoming Bias*, but Bayes nets

can be considered as an archetypal example of "architecture" and "deep insight." Also, ask yourself how lawful intelligence seemed to you before you started reading this blog, how lawful it seems to you now, then extrapolate outward from that.

* * *

Robin Hanson

I understand there are various levels on which one can express one's loves. One can love Suzy, or kind pretty funny women, or the woman selected by a panel of judges, or the the one selected by a judging process designed by a certain AI strategy, etc. But even very meta loves are loves. You want an AI that loves the choices made by a certain meta process that considers the wants of many, and that may well be a superior love. But it is still a love, your love, and the love you want to give the AI. You might think the world should be grateful to be placed under the control of such a superior love, but many of them will not see it that way; they will see your attempt to create an AI to take over the world as an act of war against them.

Eliezer Yudkowsky

Robin, using the word "love" sounds to me distinctly like something intended to evoke object-level valuation. "Love" is an archetype of direct valuation, not an archetype of metaethics.

And I'm not so much of a mutant that, rather than liking cookies, I like everyone having their reflective equilibria implemented. Taking that step is *the substance of my attempt to be fair*. In the same way that someone voluntarily splitting up a pie into three shares is not on the same moral level as someone who seizes the whole pie for themselves—even if, *by volunteering to do the fair thing rather than some other thing*, they have shown themselves to value fairness.

My take on this was given in "The Bedrock of Fairness".[2]

But you might as well say, "George Washington gave in to his desire to be a tyrant; he was just a tyrant who wanted democracy." Or, "Martin Luther King declared total war on the rest of the US, since what he wanted was a nonviolent resolution."

Similarly with "I choose not to control you" being a form of controlling.

Robin Hanson

In a foom that took two years, if the AI was visible after one year, that might give the world a year to destroy it.

Eliezer Yudkowsky

Robin, we're still talking about a local foom. Keeping security for two years may be difficult but is hardly unheard-of.

See original post for all comments.

* * *

1. Eliezer Yudkowsky, *Coherent Extrapolated Volition* (The Singularity Institute, San Francisco, CA, May 2004), http://intelligence.org/files/CEV.pdf.

2. Eliezer Yudkowsky, "The Bedrock of Fairness," *Less Wrong* (blog), July 3, 2008, http://lesswrong.com/lw/ru/the_bedrock_of_fairness/.

55

Not Taking Over the World

Eliezer Yudkowsky
15 December 2008

Followup to: What I think, If Not Why

My esteemed co-blogger Robin Hanson accuses me of *trying to take over the world*.

Why, oh why must I be so misunderstood?

(Well, it's not like I don't *enjoy* certain misunderstandings. Ah, I remember the first time someone seriously and not in a joking way accused me of trying to take over the world. On that day I felt like a true mad scientist, though I lacked a castle and hunchbacked assistant.)

But if you're working from the premise of a hard takeoff—an Artificial Intelligence that self-improves at an extremely rapid rate—and you suppose such extra-ordinary depth of insight and precision of

craftsmanship that you can *actually* specify the AI's goal system instead of *automatically* failing—

—then it takes some work to come up with a way *not* to take over the world.

Robin talks up the drama inherent in the intelligence explosion, presumably because he feels that this is a primary source of bias. But I've got to say that Robin's dramatic story does *not* sound like the story I tell of myself. There, the drama comes from tampering with such *extreme* forces that *every single idea you invent is wrong*. The standardized Final Apocalyptic Battle of Good Vs. Evil would be trivial by comparison; then all you have to do is put forth a desperate effort. Facing an adult problem in a neutral universe isn't so straightforward. Your enemy is yourself, who will *automatically* destroy the world, or just fail to accomplish anything, unless you can defeat you: That is the drama I crafted into the story I tell myself, for I too would disdain anything so cliched as Armageddon.

So, Robin, I'll ask you something of a probing question. Let's say that someone walks up to you and grants you unlimited power.

What do you do with it, so as to *not* take over the world?

Do you say, "I will do nothing—I take the null action"?

But then you have instantly become a malevolent God, as Epicurus said:

> Is God willing to prevent evil, but not able? Then he is not omnipotent.
> Is he able, but not willing? Then he is malevolent.
> Is he both able and willing? Then whence cometh evil?
> Is he neither able nor willing? Then why call him God?[1]

Peter Norvig said, "Refusing to bet is like refusing to allow time to pass."[2] The null action is also a choice. So have you not, in refusing to act, established all sick people as sick, established all poor people as poor, ordained all in despair to continue in despair, and condemned the dying to death? Will you not be, until the end of time, responsible for every sin committed?

Well, yes and no. If someone says, "I don't trust myself not to destroy the world, therefore I take the null action," then I would tend to sigh and say, "If that is so, then you did the right thing." Afterward, murderers will still be responsible for their murders, and altruists will still be creditable for the help they give.

And to say that you used your power to *take over the world* by *doing nothing to it* seems to stretch the ordinary meaning of the phrase.

But it wouldn't be the *best* thing you could do with unlimited power, either.

With "unlimited power" you have no need to crush your enemies. You have no moral defense if you treat your enemies with less than the utmost consideration.

With "unlimited power" you cannot plead the necessity of monitoring or restraining others so that they do not rebel against you. If you do such a thing, you are simply a tyrant who enjoys power, and not a defender of the people.

Unlimited power removes a lot of moral defenses, really. You can't say, "But I had to." You can't say, "Well, I wanted to help, but I couldn't." The only excuse for not helping is if you *shouldn't*, which is harder to establish.

And let us also suppose that this power is wieldable without side effects or configuration constraints; it is wielded with *unlimited precision*.

For example, you can't take refuge in saying anything like: "Well, I built this AI, but any intelligence will pursue its own interests, so now the AI will just be a Ricardian trading partner with humanity as it pursues its own goals." Say, the programming team has cracked the "hard problem of conscious experience" in sufficient depth that they can *guarantee* that the AI they create is *not sentient*—not a repository of pleasure, or pain, or subjective experience, or any interest-in-self—and hence, the AI is only a means to an end, and not an end in itself.

And you cannot take refuge in saying, "In invoking this power, the reins of destiny have passed out of my hands, and humanity has passed on the torch." Sorry, you haven't created a new person yet—not unless you *deliberately* invoke the unlimited power to do so—and then you can't take refuge in the *necessity* of it as a side effect; you must establish that it is the right thing to do.

The AI is not *necessarily* a trading partner. You could make it a nonsentient device that just gave you things, *if* you thought that were wiser.

You cannot say, "The law, in protecting the rights of all, must necessarily protect the right of Fred the Deranged to spend all day giving himself electrical shocks." The power is wielded with unlimited precision; you *could*, if you wished, protect the rights of everyone except Fred.

You cannot take refuge in the *necessity* of anything—that is the meaning of unlimited power.

We will even suppose (for it removes yet more excuses, and hence reveals more of your morality) that you are not limited by the laws of physics as we know them. You are bound to deal only in finite numbers, but not otherwise bounded. This is so that we can see the true constraints of your morality, apart from your being able to plead constraint by the environment.

In my reckless youth, I used to think that it might be a good idea to flash-upgrade to the highest possible level of intelligence you could manage on available hardware. Being smart was good, so being smarter was better, and being as smart as possible as quickly as possible was best—right?

But when I imagined having *infinite* computing power available, I realized that, no matter how large a mind you made yourself, you could just go on making yourself larger and larger and larger. So that wasn't an answer to the purpose of life. And only then did it occur to me to ask after *eudaimonic rates of intelligence increase*, rather than just assuming you wanted to immediately be as smart as possible.

Considering the infinite case moved me to change the way I considered the finite case. Before, I was *running away from the question* by saying, "More!" But considering an *unlimited* amount of ice cream forced me to confront the issue of what to *do* with *any* of it.

Similarly with population: If you invoke the unlimited power to create a quadrillion people, then why not a quintillion? If $3\uparrow\uparrow\uparrow 3$, why not $3\uparrow\uparrow\uparrow\uparrow 3$?[3] So you can't take refuge in saying, "I will create more people—that is the difficult thing, and to accomplish it is the main challenge." What is *individually* a life worth living?

You can say, "It's not my place to decide; I leave it up to others," but then you are responsible for the consequences of that decision as well. You should say, at least, how this differs from the null act.

So, Robin, reveal to us your character: What would you do with *unlimited* power?

* * *

Robin Hanson

The one ring of power sits before us on a pedestal; around it stand a dozen folks of all races. I believe that whoever grabs the ring first becomes invincible, all-powerful. If I believe we cannot make a deal, that someone is about to grab it, then I have to ask myself whether I would wield such power better than whoever I guess will grab it if I do not. If I think I'd do a better job, yes, I grab it. And I'd accept that others might consider that an act of war against them; thinking that way they may well kill me before I get to the ring.

With the ring, the first thing I do then is think very very carefully about what to do next. Most likely the first task is who to get advice from. And then I listen to that advice.

Yes, this is a very dramatic story, one which we are therefore biased to overestimate its likelihood.

I don't recall where exactly, but I'm pretty sure I've already admitted that I'd "grab the ring" before on this blog in the last month.

Eliezer Yudkowsky

I'm not asking you *if* you'll take the Ring, I'm asking *what you'll do with the Ring*. It's already been handed to you.

Take advice? That's still something of an evasion. What advice would *you* offer you? You don't seem quite satisfied with what (you think) is my plan for

the Ring—so you must *already* have an opinion of your own—what would you change?

Robin Hanson

Eliezer, I haven't meant to express any dissatisfaction with your plans to use a ring of power. And I agree that someone should be working on such plans even if the chances of it happening are rather small. So I approve of your working on such plans. My objection is only that if enough people overestimate the chance of such scenario, it will divert too much attention from other important scenarios. I similarly think global warming is real, worthy of real attention, but that it diverts too much attention from other future issues.

Eliezer Yudkowsky

Okay, you don't disapprove. Then consider the question one of curiosity. If Tyler Cowen acquired a Ring of Power and began gathering a circle of advisors, and you were in that circle, what specific advice would you give him?

Robin Hanson

Eliezer, I'd advise no sudden moves; think very carefully before doing anything. I don't know what I'd think after thinking carefully, as otherwise I wouldn't need to do it. Are you sure there isn't some way to delay thinking on your problem until after it appears? Having to have an answer now when it seems an unlikely problem is very expensive.

See original post for all comments.

* * *

1. Goodreads, "Epicurus Quotes," 2013, accessed July 28, 2013, http://www.goodreads.com/author/quotes/114041.Epicurus.

2. Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 1st ed. (Upper Saddle River, NJ: Prentice-Hall, 1995).

3. See http://en.wikipedia.org/wiki/Knuth%27s_up-arrow_notation for an explanation of this notation for very large numbers.

Part IV

Postscript

56

We Agree: Get Froze

Robin Hanson

12 December 2008

My co-blogger Eliezer and I may disagree on AI fooms, but we agree on something quite contrarian and, we think, huge: ***More likely than not, most folks who die today didn't have to die!*** Yes, I am skeptical of most medicine because on average it seems folks who get more medicine aren't healthier.[1] But I'll heartily endorse one medical procedure: *cryonics*, i.e., freezing folks in liquid nitrogen when the rest of medicine gives up on them.

Yes, even with modern antifreezes, freezing does lots of damage, perhaps more than whatever else was going to kill you. But bodies frozen that cold basically won't change for millennia. So if whole-brain emulation is ever achieved, and if freezing doesn't destroy info needed for an em scan, if we think more likely than not future folks could make an em out of your frozen brain. Since most folks who

die today have an intact brain until the rest of their body fails them, more likely than not most death victims today could live on as (one or more) future ems. And if future folks learn to repair freezing damage plus whatever was killing victims, victims might live on as ordinary humans.

Now there are a few complications:

- *If too many folks are frozen, the future might not want to revive them all.* But in four decades of cryonics, only about a thousand folks have signed up, and a hundred have actually been frozen.[2] So this isn't remotely problem yet. And by investing, frozen folk could easy *pay* to be revived.

- *Some people don't want to live as future ems.* Maybe we'll just have to let such prudes die.

- *Many people don't want to come back to a world without their friends and associates.* But the more who are frozen, the less of a problem this becomes. Sign up *together* with your loved ones.

- *Organizations charged with keeping bodies frozen could fail before revival is possible.* But the more who are frozen, the less often this will happen, and the cheaper cryonics will become as well. There are huge scale economies to freezing folks.

Amazingly, while we subsidize most medicine but gain little directly from that, we actively discourage cryonics, which could literally save billions of lives. No health insurance covers it, it gets no government subsidy, doctors won't call it "medicine," and it has to be done under the fiction of "organ donation," as frozen folks are legally "dead." And

in a society that is relatively tolerant of various religious beliefs and funeral procedures, prosecutors often attack it, family members often actively prevent relatives from being frozen, and spouses commonly threaten to divorce folks wanting to be frozen.[3] (HT to Kerry Howley.)

It seems far more people read this blog daily than have ever signed up for cryonics. While it is hard to justify most medical procedures using standard health economics calculations, such calculations say that at today's prices cryonics seems a good deal even if you think there's only a 5% chance it'll work—at least if you have a typical US income and think you'd enjoy living in a future world. In addition, you'd make it easier for others to avoid death. It really is hard to find a clearer example of an avoidable Holocaust that you can personally do something substantial about now. And you'd help yourself in the process!

If anyone here disagrees, do speak up, as should any influential blogger out there who wants to debate this. You who agree, however, let other readers here know it isn't just the two of us. The rest of you, consider saving your life!

* * *

See original post for all comments.

* * *

1. Robin Hanson, "Cut Medicine In Half," *Overcoming Bias* (blog), September 10, 2007, http://www.overcomingbias.com/2007/09/cut-medicine-in.html.

2. Alcor Life Extension Foundation, "Alcor Membership Statistics," April 30, 2013, accessed July 28, 2013, http://www.alcor.org/AboutAlcor/membershipstats.html.

3. Michael G. Darwin, Chana de Wolf, and Aschwin de Wolf, "Is That What Love Is? The Hostile Wife Phenomenon in Cryonics," *Evidence Based Cryonics* (blog), 2008, http://www.evidencebasedcryonics.org/is-that-what-love-is-the-hostile-wife-phenomenon-in-cryonics/.

57

You Only Live Twice

Eliezer Yudkowsky
12 December 2008

> It just so happens that your friend here is only *mostly* dead. There's a big difference between *mostly* dead and *all* dead.
>
> —*The Princess Bride*[1]

My co-blogger Robin and I may disagree on how fast an AI can improve itself, but we agree on an issue that seems much simpler to us than that: **At the point where the current legal and medical system gives up on a patient, they aren't really dead.**

Robin has already said much of what needs saying, but a few more points:

- Ben Best's Cryonics FAQ,[2] Alcor's FAQ,[3] Alcor FAQ for scientists,[4] Scientists' Open Letter on Cryonics[5]

- I know more people who are planning to sign up for cryonics Real Soon Now than people who have actually signed up. **I expect that more people have *died while cryocrastinating than have actually been cryopreserved*.** If you've *already decided* this is a good idea, but you "haven't gotten around to it," sign up for cryonics NOW. I mean RIGHT NOW. Go to the website of Alcor or the Cryonics Institute and follow the instructions.

- Cryonics is usually funded through life insurance. The following conversation from an Overcoming Bias meetup is worth quoting:

> HIM: I've been thinking about signing up for cryonics when I've got enough money.
>
> ME: Um . . . it doesn't take all that much money.
>
> HIM: It doesn't?
>
> ME: Alcor is the high-priced high-quality organization, which is something like $500–$1,000 in annual fees for the organization, I'm not sure how much. I'm young, so I'm signed up with the Cryonics Institute, which is $120/year for the membership. I pay $180/year for more insurance than I need—it'd be enough for Alcor too.
>
> HIM: That's ridiculous.
>
> ME: Yes.
>
> HIM: No, really, that's *ridiculous*. If that's true then my decision isn't just determined, it's overdetermined.
>
> ME: Yes. And there's around a thousand people worldwide [actually 1,400] who are signed up for cryonics. Figure that at most a quarter of those did it for systematically rational reasons. That's a high upper bound on the number of people on Earth who can reliably reach the right conclusion on massively overdetermined issues.

- Cryonics is not marketed well—or at all, really. There's no salespeople who get commissions. There is *no one to hold your hand through signing up*, so you're going to have to get the papers signed and notarized yourself. The closest thing out there might be Rudi Hoffman, who sells life insurance with cryonics-friendly insurance providers (I went through him).

- If you want to *securely* erase a hard drive, it's not as easy as writing it over with zeroes. Sure, an "erased" hard drive like this won't boot up your computer if you just plug it in again. But if the drive falls into the hands of a specialist with a scanning tunneling microscope, they can tell the difference between "this was a 0, overwritten by a 0" and "this was a 1, overwritten by a 0."

  There are programs advertised to "securely erase" hard drives using many overwrites of 0s, 1s, and random data. But if you want to keep the secret on your hard drive secure against *all possible future technologies that might ever be developed*, then cover it with thermite and set it on fire. It's the only way to be sure.

  Pumping someone full of cryoprotectant and gradually lowering their temperature until they can be stored in liquid nitrogen is not a secure way to erase a person.

  See also the information-theoretic criterion of death (Wikipedia).

- You don't have to buy what's usually called the "patternist" philosophy of identity to sign up for cryonics. After reading all the

information off the brain, you could put the "same atoms" back into their old places.

- "Same atoms" is in scare quotes because our current physics *prohibits* particles from possessing individual identities. It's a much stronger statement than "we can't tell the particles apart with current measurements" and has to do with the notion of configuration spaces in quantum mechanics. This is a standard idea in QM, *not* an unusual woo-woo one—see the Quantum Physics sequence on *Less Wrong* for a gentle introduction. Although patternism is not *necessary* to the cryonics thesis, we happen to live in a universe where "the same atoms" is physical nonsense.

There's a number of intuitions we have in our brains for processing a world of distinct physical objects, built in from a very young age. These intuitions, which may say things like, "If an object disappears, and then comes back, it isn't the same object," are tuned to our macroscopic world and generally don't match up well with *fundamental* physics. Your identity is not like a little billiard ball that follows you around—there aren't *actually* any billiard balls down there.

Separately and convergently, more abstract reasoning strongly suggests that "identity" should not be epiphenomenal; that is, you should not be able to change someone's identity without changing any observable fact about them.

If you go through the aforementioned *Less Wrong* sequence, you should actually be able to *see intuitively* that successful cryonics preserves anything about you that is preserved by going to sleep at night and waking up the next morning.

Cryonics, to me, makes two statements.

The first statement is about systematically valuing human life. It's bad when a pretty young white girl goes missing somewhere in America. But when 800,000 Africans get murdered in Rwanda, that gets 1/134 the media coverage of the Michael Jackson trial. It's sad, to be sure, but no cause for emotional alarm. When brown people die, that's *all part of the plan*—as a smiling man once said.

Cryonicists are people who've decided that their deaths, and the deaths of their friends and family and the rest of the human species, are not part of the plan.[6]

I've met one or two Randian-type "selfish" cryonicists, but they aren't a majority. Most people who sign up for cryonics wish that everyone would sign up for cryonics.

The second statement is that you have at least a *little* hope in the future. Not faith, not blind hope, not irrational hope—just any hope at all.

I was once at a table with Ralph Merkle, talking about how to market cryonics if anyone ever gets around to marketing it, and Ralph suggested a group of people in a restaurant, having a party; and the camera pulls back, and moves outside the window, and the restaurant is on the Moon. Tagline: "Wouldn't you want to be there?"

If you look back at, say, the Middle Ages, things were worse then. I'd rather live here then there. I have hope that humanity will move forward *further*, and that's something that I want to see.

And I hope that the idea that people are disposable, and that their deaths are part of the plan, is something that fades out of the Future.

Once upon a time, infant deaths were part of the plan, and now they're not. Once upon a time, slavery was part of the plan, and now it's not. Once upon a time, dying at thirty was part of the plan, and now it's not. That's a psychological shift, not just an increase in living standards. Our era doesn't value human life with perfect consistency—but the value of human life is higher than it once was.

We have a concept of what a medieval peasant *should* have had, the dignity with which they *should* have been treated, that is higher than what they would have thought to ask for themselves.

If no one in the future cares enough to save people who can be saved . . . well. In cryonics there is an element of taking responsibility for the Future. You may be around to reap what your era has sown. It is not just my *hope* that the Future be a better place; it is my *responsibility*. If I thought that we were on track to a Future where no one cares about human life, and lives that could easily be saved are just thrown away—then I would try to change that. Not everything worth doing is easy.

Not signing up for cryonics—what does that say? That you've lost hope in the future. That you've lost your will to live. That you've stopped believing that human life, and your own life, is something of value.

This can be a painful world we live in, and the media is always telling us how much worse it will get. If you spend enough time not looking forward to the next day, it damages you, after a while. You lose your ability to hope. Try telling someone already grown old to sign up for cryonics, and they'll tell you that they don't want to be old forever—that they're tired. If you try to explain to someone already grown old, that the nanotechnology to revive a cryonics patient

is sufficiently advanced that reversing aging is almost trivial by comparison . . . then it's not something they can imagine on an emotional level, no matter what they believe or don't believe about future technology. They can't imagine not being tired. I think that's true of a lot of people in this world. If you've been hurt enough, you can no longer imagine healing.

But things really were a lot worse in the Middle Ages. And they really are a lot better now. Maybe humanity *isn't* doomed. The Future could be something that's worth seeing, worth living in. And it may have a concept of sentient dignity that values your life more than you dare to value yourself.

On behalf of the Future, then—please ask for a little more for yourself. More than death. It really . . . isn't being selfish. *I* want you to live. I think that the Future will want you to live. That if you let yourself die, people who aren't even born yet will be sad for the irreplaceable thing that was lost.

So please, live.

My brother didn't. My grandparents won't. But everything we can hold back from the Reaper, even a single life, is precious.

If other people want you to live, then it's not just you doing something selfish and unforgivable, right?

So I'm saying it to you.

I want you to live.

* * *

Robin Hanson

Eliezer, well written! :)

See original post for all comments.

* * *

1. Rob Reiner, dir., *The Princess Bride*, prod. Andrew Scheinman, **cbywritten by**William Goldman (20th Century Fox, September 25, 1987), film.

2. Ben Best, "Cryonics — Frequently Asked Questions (FAQ)," 2004, last revised August 22, 2012, http://www.benbest.com/cryonics/CryoFAQ.html.

3. Alcor Life Extension Foundation, "Frequently Asked Questions," accessed July 28, 2013, http://www.alcor.org/FAQs/index.html.

4. Alcor Life Extension Foundation, "Scientists' Cryonics FAQ," accessed July 28, 2013, http://www.alcor.org/sciencefaq.htm.

5. Gregory Benford et al., "Scientists' Open Letter on Cryonics," accessed July 24, 2013, http://www.evidencebasedcryonics.org/scientists-open-letter-on-cryonics/.

6. Eliezer Yudkowsky, "Yehuda Yudkowsky, 1985–2004," November 2004, last revised May 8, 2005, http://yudkowsky.net/other/yehuda.

58

Hanson-Yudkowsky Jane Street Debate 2011

Robin Hanson and Eliezer Yudkowsky

29 June 2011

Moderator: Do you want to say what the statement is?

Eliezer Yudkowsky: I forget what the exact form of it was. The question is, "After all sorts of interesting technological things happen at some undetermined point in the future, are we going to see a very small nucleus that can or does control all the resources, or do we see a general, more civilization-wide, large fraction of society participating in all these things going down?"

Robin Hanson: I think, if I remember it, it was, "Compared to the industrial and farming revolutions, intelligence-explosion first movers will soon dominate a larger fraction of the future world."

Eliezer: That's what I remember.

Moderator: There was a whole debate to get to this statement.

(*Laughter.*)

MODERATOR: Right, so, "for"—

ROBIN: We'll try to explain what those mean.

MODERATOR: "For" is saying that you believe that the first movers will gain a large lead relative to first movers in the industrial and farming revolutions.

ROBIN: Right.

MODERATOR: If you agree with that statement, you're "for."

ROBIN: This side. (*Gestures to Eliezer.*)

MODERATOR: If you think it's going to be more broad-based . . .

ROBIN: Con. (*Gestures toward self.*)

ELIEZER: Maybe a one-word thing would be "highly centralized," "highly decentralized." Does that sound like a one-word—?

ROBIN: There has to be a cutoff in between "highly," so (*laughs*) there's that middle ground.

ELIEZER: With the cutoff point being the agricultural revolution, for example. Or no, that's actually not the cutoff point. That's your side.

MODERATOR: On the yellow sheet, if you're in favor, you write your name and "I'm in favor." If you're against, you write your name and "I'm against." Then pass them that way. Keep the colored sheet; that's going to be your vote afterwards. Eliezer and Robin are hoping to convert you.

ROBIN: Or have fun.

MODERATOR: What?

ROBIN: Or have fun trying.

MODERATOR: We're very excited at Jane Street today to have Eliezer Yudkowsky, Robin Hanson.

(*Applause.*)

MODERATOR: I'll keep the intros short so we can jump into the debate. Both very highly regarded intellectuals and have been airing this debate for some time, so it should be a lot of fun.

(*Gestures to Robin Hanson.*) Professor at George Mason University of economics, one of the frontiers in prediction markets, all the way back to 1988. Avid publisher. Both a cofounder of *Overcoming Bias*, now he's moved over to *Less Wrong*.

ELIEZER: Oh, I moved over to *Less Wrong*, and he's at *Overcoming Bias*.

MODERATOR: Eliezer, a cofounder of the Singularity Institute. Many, many publications. Without further ado, on to the debate, and . . . first five minutes.

(*Laughter.*)

ELIEZER: Quick question. How many people here are already familiar with the differences between what Ray Kurzweil means when he uses the word "singularity" and what the Singularity Institute means when they use the word "singularity"? Raise your hand if you're already familiar with the difference. OK. I don't see a sea of hands. That means that I designed this talk correctly.

You've probably run across a word, "singularity." People use it with a lot of different and mutually incompatible meanings. When we named the Singularity Institute for Artificial Intelligence in 2000, it meant something pretty different then than now.

The original meaning was—a mathematician and science fiction writer named Vernor Vinge originally coined the word "singularity" to describe the breakdown in his ability to model and imagine the future when he tried to extrapolate that model past the point where it

predicted the technological creation of smarter-than-human intelligence. In this particular case, he was trying to write a story about a human with a brain-computer interface increasing his intelligence. The rejection letter he got from John Campbell said, "Sorry—you can't write this story. Neither can anyone else."

If you asked an ancient Greek from 2,500 years ago to imagine the modern world, in point of fact they wouldn't be able to, but they'd have much better luck imagining our world and would manage to get more things right than, say, a chimpanzee would. There are stories from thousands of years ago that still resonate with us today, because the minds, the brains haven't really changed over that time. If you change the brain, the mind, that implies a difference in the future that is different in kind from faster cars or interplanetary travel or curing cancer or bionic arms or similar such neat, cool technological trivia, because that would not really have an impact on the future comparable to the rise of human intelligence fifty thousand years ago.

The other thing is that since intelligence is the source of technology—that is, *this* is ultimately the factor that produces the chairs, the floor, the projectors, this computer in front of me—if you tamper with this, then you would expect that to ripple down the causal chain and, in other words, if you make this more powerful, you get a different kind of technological impact than you get from any one breakthrough.

I. J. Good, another mathematician, coined a related concept of the singularity when he pointed out that if you could build an artificial intelligence that was smarter than you, it would also be better than you at designing and programming artificial intelligence. So this AI builds an even smarter AI, or instead of a whole other AI, just re-

programs modules within itself, then that AI build an even smarter one . . .

I. J. Good suggested that you'd get a positive feedback loop leading to what I. J. Good termed "ultraintelligence" but what is now generally called "superintelligence," and the general phenomenon of smarter minds building even smarter minds is what I. J. Good termed the "intelligence explosion."

You could get an intelligence explosion outside of AI. For example, humans with brain-computer interfaces designing the next generation of brain-computer interfaces. But the purest and fastest form of the intelligence explosion seems likely to be an AI rewriting its own source code.

This is what the Singularity Institute is actually about. If we'd foreseen what the word "singularity" was going to turn into, we'd have called ourselves the "Good Institute" or the "Institute for Carefully Programmed Intelligence Explosions."

(*Laughter.*)

Eliezer: Here at the Institute for Carefully Programmed Intelligence Explosions, we do not necessarily believe or advocate that, for example, there was more change in the forty years between 1970 and 2010 than the forty years between 1930 and 1970.

I myself do not have a strong opinion that I could argue on this subject, but our president Michael Vassar, our major donor Peter Thiel, and Thiel's friend Kasparov, who, I believe, recently spoke here, all believe that it's obviously wrong that technological change has been accelerating at all, let alone that it's been accelerating exponentially. This doesn't contradict the basic thesis that we would advocate, because you do not need exponentially accelerating technologi-

cal progress to eventually get an AI. You just need some form of technological progress, period.

When we try to visualize how all this is likely to go down, we tend to visualize a scenario that someone else once termed "a brain in a box in a basement." I love that phrase, so I stole it. In other words, we tend to visualize that there's this AI programming team, a lot like the sort of wannabe AI programming teams you see nowadays, trying to create artificial general intelligence, like the artificial general intelligence projects you see nowadays. They manage to acquire some new deep insights which, combined with published insights in the general scientific community, let them go down into their basement and work in it for a while and create an AI which is smart enough to reprogram itself, and then you get an intelligence explosion.

One of the strongest critics of this particular concept of a localized intelligence explosion is Robin Hanson. In fact, it's probably fair to say that he is the strongest critic by around an order of magnitude and a margin so large that there's no obvious second contender.

(*Laughter.*)

Eliezer: How much time do I have left in my five minutes? Does anyone know, or . . . ?

Moderator: You just hit five minutes, but—

Eliezer: All right. In that case, I'll turn you over to Robin.

(*Laughter.*)

Robin: We're going to be very flexible here going back and forth, so there'll be plenty of time. I thank you for inviting us. I greatly respect this audience and my esteemed debate opponent here. We've known each other for a long time. We respect each other, we've talked a lot. It's a lot of fun to talk about this here with you all.

The key question here, as we agree, is this idea of a local intelligence explosion. That's what the topic's about. We're not talking about this idea of gradually accelerating change, where in thirty years everything you've ever heard about will all be true or more. We're talking about a world where we've had relatively steady change over a century, roughly, and we might have steady change for a while, and then the hypothesis is there'll be this sudden dramatic event with great consequences, and the issue is, what is the nature of that event, and how will it play out?

This "brain in a box in a basement" scenario is where something that starts out very small, very quickly becomes very big. And the way it goes from being small to being very big is it gets better. It gets more powerful. So, in essence, during this time this thing in the basement is outcompeting the entire rest of the world.

Now, as you know, or maybe you don't know, the world today is vastly more powerful than it has been in the past. The long-term history of your civilization, your species, has been a vast increase in capacity. From primates to humans with language, eventually developing farming, then industry, and who knows where, over this very long time, lots and lots of things have been developed, lots of innovations have happened.

There's lots of big stories along the line, but the major, overall, standing-from-a-distance story is of relatively steady, gradual growth. That is, there's lots of inventions here, changes there, that add up to disruptions, but most of the disruptions are relatively small and on the distant scale there's relatively steady growth. It's more steady, even, on the larger scales. If you look at a company like yours, or a city, even,

like this, you'll have ups and downs, or even a country, but on the long timescale . . .

This is central to the idea of where innovation comes from, and that's the center of this debate, really. Where does innovation come from, where can it come from, and how fast can it come?

So the brain in the box in the basement—within a relatively short time a huge amount of innovation happens, that is this thing hardly knows anything, it's hardly able to do anything, and then within a short time it's able to do so much that it basically can take over the world and do whatever it wants, and that's the problem.

Now let me stipulate right from the front, there is a chance he's right. OK? And somebody ought to be working on that chance. He looks like a good candidate to me, so I'm fine with him working on this chance. I'm fine with there being a bunch of people working on the chance. My only dispute is the perceptions of probability. Some people seem to think this is the main, most likely thing that's going to happen. I think it's a small chance that's worth looking into and protecting against, so we all agree there. Our dispute is more about the chance of this scenario.

If you remember the old Bond villain, he had an island somewhere with jumpsuited minions, all wearing the same color if I recall. They had some device they invented, and Bond had to go in and put it off. Usually, they had invented a whole bunch of devices back there, and they just had a whole bunch of stuff going on.

Sort of the epitome of this might be Captain Nemo, from *Twenty Thousand Leagues Under the Sea*. One guy off on his own island with a couple of people invented the entire submarine technology, if you

believe the movie, undersea cities, nuclear weapons, etc., all within a short time.

Now, that makes wonderful fiction. You'd like to have a great powerful villain that everybody can go fight and take down. But in the real world it's very hard to imagine somebody isolated on an island with a few people inventing large amounts of technology, innovating, and competing with the rest of the world.

That's just not going to happen, it doesn't happen in the real world. In our world, so far, in history, it's been very rare for any one local place to have such an advantage in technology that it really could do anything remotely like take over the world.

In fact, if we look for major disruptions in history, which might be parallel to what's being hypothesized here, the three major disruptions you might think about would be the introductions of something special about humans (perhaps language), the introduction of farming, and the introduction of industry.

Those three events—whatever was special about them we're not sure, but for those three events the growth rate of the world economy suddenly, within a very short time, changed from something that was slow to something a hundred or more times faster. We're not sure exactly what those were, but those would be candidates, things I would call singularities, that is big, enormous disruptions.

But in those singularities, the places that first had the new technology had varying degrees of how much an advantage they gave. Edinburgh gained some advantage by being the beginning of the Industrial Revolution, but it didn't take over the world. Northern Europe did more like take over the world, but even then it's not so much taken over the world. Edinburgh and parts of Northern Europe needed each

other. They needed a large economy to build things together, so that limited . . . Also, people could copy. Even in the farming revolution, it was more like a fifty–fifty split between the initial farmers spreading out and taking over territory and the other locals copying them and interbreeding with them.

If you go all the way back to the introduction of humans, that was much more about one displaces all the rest because there was relatively little way in which they could help each other, complement each other, or share technology.

What the issue here is—and obviously I'm done with my five minutes—in this new imagined scenario, how plausible is it that something that's very small could have that much of an advantage? That whatever it has that's new and better gives it such an advantage that it can grow from something that's small, on even a town scale, to being bigger than the world, when it's competing against the entire rest of the world? When, in these previous innovation situations where even the most disruptive things that ever happened, still, the new first mover only gained a modest advantage in terms of being a larger fraction of the new world.

I'll end my five minutes there.

Eliezer: The fundamental question of rationality is, what do you think you know and how you do think you know it? This is rather interesting and in fact, it's rather embarrassing, because it seems to me like there's very strong reason to believe that we're going to be looking at a localized intelligence explosion.

Robin Hanson feels there's pretty strong reason to believe that we're going to be looking at a nonlocal general economic growth mode changeover. Calling it a singularity seems . . . Putting them all

into the category of singularity is a slightly begging the definitional question. I would prefer to talk about the intelligence explosion as a possible candidate for the reference class "economic growth mode changeovers."

Robin: OK.

Eliezer: The embarrassing part is that both of us know the theorem which shows that two rational agents cannot agree to have common knowledge of disagreement, called Aumann's Agreement Theorem. So we're supposed to, since we know that the other person believes something different, we're supposed to have agreed by now, but we haven't. It's really quite embarrassing.

But the underlying question is, is the next big thing going to look more like the rise of human intelligence, or is it going to look more like the Industrial Revolution? If you look at modern AI projects, the leading edge of artificial intelligence does not look like the product of an economy among AI projects.

They tend to rewrite their own code. They tend to not use very much cognitive content that other AI projects have developed. They've been known to import libraries that have been published, but you couldn't look at that and say that an AI project which just used what had been published, and then developed its own further code, would suffer a disadvantage analogous to a country that tried to go its own way for the rest of the world economy.

Rather, AI projects nowadays look a lot like species, which only share genes within a species and then the other species are all off going their own way.

(*Gestures to Robin.*) What is your vision of the development of intelligence or technology where things are getting traded very quickly, analogous to the global economy?

Robin: Let's back up and make sure we aren't losing people with some common terminology. I believe, like most of you do, that in the near future, within a century, we will move more of the knowledge and intelligence in our society into machines. That is, machines have a lot of promise as hardware substrate for intelligence. You can copy them. You can reproduce them. You can make them go faster. You can have them in environments. We are in complete agreement that eventually hardware, nonbiological hardware, silicon, things like that, will be a more dominant substrate of where intelligence resides. By intelligence, I just mean whatever mental capacities exist that allow us to do mental tasks.

We are a powerful civilization able to do many mental tasks, primarily because we rely heavily on bodies like yours with heads like yours where a lot of that stuff happens inside—biological heads. But we agree that in the future there will be much more of that happening in machines. The question is the path to that situation.

Now, our heritage, what we have as a civilization, a lot of it is the things inside people's heads. Part of it isn't what was in people's heads fifty thousand years ago. But a lot of it is also just what was in people's heads fifty thousand years ago. We have this common heritage of brains and minds that goes back millions of years to animals and built up with humans and that's part of our common heritage.

There's a lot in there. Human brains contain an enormous amount of things. I think it's not just one or two clever algorithms or something, it's this vast pool of resources. It's like comparing it to a city,

like New York City. New York City is a vast, powerful thing because it has lots and lots of stuff in it.

When you think in the future there will be these machines and they will have a lot of intelligence in them, one of the key questions is, "Where will all of this vast mental capacity that's inside them come from?" Where Eliezer and I differ, I think, is that I think we all have this vast capacity in our heads and these machines are just way, way behind us at the moment, and basically they have to somehow get what's in our head transferred over to them somehow. Because if you just put one box in a basement and ask it to rediscover the entire world, it's just way behind us. Unless it has some almost inconceivable advantage over us at learning and growing and discovering things for itself, it's just going to remain way behind unless there's some way it can inherit what we have.

ELIEZER: OK. I gave a talk here at Jane Street that was on the speed of evolution. Raise your hand if you were here for this and remember some of it. OK.

(*Laughter.*)

ELIEZER: There's a single, simple algorithm which produced the design for the human brain. It's not a very good algorithm, it's extremely slow. It took it millions and millions and billions of years to cough up this artifact over here. (*Gestures to head.*) Evolution is so simple and so slow that we can even make mathematical statements about how slow it is, such as the two separate bounds that I've seen calculated for how fast evolution can work, one of which is on the order of one bit per generation, in the sense that, let's say, two parents have sixteen children, then on average, all but two of those children must die or fail to reproduce or the population goes to zero or infinity

very rapidly. Sixteen cut down to two, that would be three bits of selection pressure per generation. There's another argument which says that it's faster than this.

But if you actually look at the genome, then we've got about thirty thousand genes in here, most of our 750 megabytes of DNA is repetitive and almost certainly junk, as best we understand it, and the brain is simply not a very complicated artifact by comparison to, say, Windows Vista. Now, the complexity that it does have, it uses a lot more effectively than Windows Vista does. It probably contains a number of design principles which Microsoft knows not.

But nonetheless, what I'm trying to say is . . . I'm not saying that it's that small because it's 750 megabytes, I'm saying it's got to be that small because most of it, at least 90% of the 750 megabytes is junk and there's only thirty thousand genes for the whole body, never mind the brain.

That something that simple can be this powerful and this hard to understand is a shock. But if you look at the brain design, it's got fifty-two major areas on each side of the cerebral cortex, distinguishable by the local pattern, the tiles and so on. It just doesn't really look all that complicated. It's very powerful. It's very mysterious. What we can't say about it is that it probably involves one thousand different deep major mathematical insights into the nature of intelligence that we need to comprehend before we can build it.

This is probably one of the more intuitive, less easily quantified, and argued by reference to large bodies of experimental evidence type things. It's more a sense of, well, you read through *The MIT Encyclopedia of Cognitive Sciences*, and you read Judea Pearl's *Probabilistic Reasoning in Intelligent Systems*. Here's an insight. It's an insight into

the nature of causality. How many more insights of this size do we need, given that this is what the *The MIT Encyclopedia of Cognitive Sciences* seems to indicate we already understand and this is what we don't? You take a gander at it, and you say there's probably about ten more insights. Definitely not one. Not a thousand. Probably not a hundred either.

Robin: To clarify what's at issue: The question is, what makes your human brain powerful?

Most people who look at the brain and compare it to other known systems have said things like "It's the most complicated system we know," or things like that. Automobiles are also powerful things, but they're vastly simpler than the human brain, at least in terms of the fundamental constructs.

But the question is, what makes the brain powerful? Because we won't have a machine that competes with the brain until we have it have whatever the brain has that makes it so good. So the key question is, what makes the brain so good?

I think our dispute in part comes down to an inclination toward architecture or content. That is, one view is that there's just a clever structure and if you have that basic structure, you have the right sort of architecture, and you set it up that way, then you don't need very much else. You just give it some sense organs, some access to the Internet or something, and then it can grow and build itself up because it has the right architecture for growth. Here we mean architecture for growth in particular—what architecture will let this thing grow well?

Eliezer hypothesizes that there are these insights out there, and you need to find them. And when you find enough of them, then

you can have something that competes well with the brain at growing because you have enough of these architectural insights.

My opinion, which I think many AI experts will agree with at least, including say Doug Lenat, who did the Eurisko program that you (*gesturing toward Eliezer*) most admire in AI, is that it's largely about content. There are architectural insights. There are high-level things that you can do right or wrong, but they don't, in the end, add up to enough to make vast growth. What you need for vast growth is simply to have a big base.

In the world, there are all these nations. Some are small. Some are large. Large nations can grow larger because they start out large. Cities like New York City can grow larger because they start out as larger cities.

If you took a city like New York and you said, "New York's a decent city. It's all right. But look at all these architectural failings. Look how this is designed badly or that's designed badly. The roads are in the wrong place or the subways are in the wrong place or the building heights are wrong, or the pipe format is wrong. Let's imagine building a whole new city somewhere with the right sort of architecture." How good would that better architecture have to be?

You clear out some spot in the desert. You have a new architecture. You say, "Come, world, we have a better architecture here. You don't want those old cities. You want our new, better city." I predict you won't get many comers because, for cities, architecture matters, but it's not that important. It's just lots of people being there and doing lots of specific things that makes a city better.

Similarly, I think that for a mind, what matters is that it just has lots of good, powerful stuff in it, lots of things it knows, routines, strategies, and there isn't that much at the large architectural level.

Eliezer: The fundamental thing about our modern civilization is that everything you've ever met that you bothered to regard as any sort of ally or competitor had essentially exactly the same architecture as you.

In the logic of evolution in a sexually reproducing species, you can't have half the people having a complex machine that requires ten genes to build, because then if all the individual genes are at 50% frequency, the whole thing only gets assembled 0.1% of the time. Everything evolves piece by piece, piecemeal. This, by the way, is standard evolutionary biology. It's not a creationist argument. I just thought I would emphasize that in case anyone was . . . This is bog standard evolutionary biology.

Everyone you've met, unless they've suffered specific brain damage or a specific genetic deficit, they have all the same machinery as you. They have no complex machine in their brain that you do not have.

Our nearest neighbors, the chimpanzees, who have 95% shared DNA with us . . . Now, in one sense, that may be a little misleading because what they don't share is probably more heavily focused on brain than body type stuff, but on the other hand, you can look at those brains. You can put the brains through an MRI. They have almost exactly the same brain areas as us. We just have larger versions of some brain areas. I think there's one sort of neuron that we have and they don't, or possibly even they had it but only in very tiny quantities.

This is because there have been only five million years since we split off from the chimpanzees. There simply has not been time to do any major changes to brain architecture in five million years. It's just not enough to do really significant complex machinery. The intelligence we have is the last layer of icing on the cake and yet, if you look at the sort of curve of evolutionary optimization into the hominid line versus how much optimization power is put out, how much horsepower was the intelligence, it goes like this. (*Gestures a flat line, then a sharp vertical increase, then another flat line.*)

If we look at the world today, we find that taking a little bit out of the architecture produces something that is just not in the running as an ally or a competitor when it comes to doing cognitive labor. Chimpanzees don't really participate in the economy at all, in fact, but the key point from our perspective is that, although they are in a different environment, they grow up learning to do different things, there are genuinely skills that chimpanzees have that we don't, such as being able to poke a branch into an anthill and draw it out in such a way as to have it covered with lots of tasty ants. Nonetheless, there are no branches of science where the chimps do better because they have mostly the same architecture and more relevant content.

It seems to me at least that if we look at the present cognitive landscape, we're getting really strong information that—you can imagine that we're trying to reason from one sample, but then pretty much all of this is reasoning from one sample in one way or another—we're seeing that in this particular case at least, humans can develop all sorts of content that lets them totally outcompete other animal species who have been doing things for millions of years longer than we have by

virtue of architecture, and anyone who doesn't have the architecture isn't really in the running for it.

Robin: So something happened to humans. I'm happy to grant that humans are outcompeting all the rest of the species on the planet.

We don't know exactly what it is about humans that was different. We don't actually know how much of it was architecture, in a sense, versus other things. But what we can say, for example, is that chimpanzees actually could do a lot of things in our society, except they aren't domesticated.

The animals we actually use are a very small fraction of the animals out there. It's not because they're smarter, *per se*, it's because they are just more willing to be told what to do. Most animals aren't willing to be told what to do. If chimps would be willing to be told what to do, there's a lot of things we could have them do. *Planet of the Apes* would actually be a much more feasible scenario. It's not clear that their cognitive abilities are really that lagging, more that their social skills are lacking.

But the more fundamental point is that, since a million years ago when humans probably had language, we are now a vastly more powerful species, because we used this ability to collect cultural content and built up a vast society that contains so much more. I think that if you took humans and made some better architectural innovations to them and put a pile of them off in the forest somewhere, we're still going to outcompete them if they're isolated from us because we just have this vaster base that we have built up since then.

Again, the issue comes down to, how important is architecture? Even if something happened such that some architectural thing finally enabled humans to have culture, to share culture, to have lan-

guage, to talk to each other, that was powerful—the question is, how many more of those are there? Because we have to hypothesize not just that there are one or two, but there are a whole bunch of these things, because that's the whole scenario, remember?

The scenario is: box in a basement, somebody writes the right sort of code, turns it on. This thing hardly knows anything, but because it has all these architectural insights, it can in a short time take over the world. There have to be a lot of really powerful architectural low-hanging fruit to find in order for that scenario to work. It's not just a few ways in which architecture helps, it's architectural dominance.

ELIEZER: I'm not sure I would agree that you need lots of architectural insights like that. I mean, to me, it seems more like you just need one or two.

ROBIN: But one architectural insight allows a box in a basement that hardly knows anything to outcompete the entire rest of the world?

ELIEZER: Well, if you look at humans, they outcompeted everything evolving, as it were, in the sense that there was this one optimization process, natural selection, that was building up content over millions and millions and millions of years, and then there's this new architecture which can all of the sudden generate vast amounts—

ROBIN: So humans can accumulate culture, but you're thinking there's another thing that's metaculture that these machines will accumulate that we aren't accumulating?

ELIEZER: I'm pointing out that the timescale for generating content underwent this vast temporal compression. In other words, content that used to take millions of years to do now can now be done on the order of hours.

Robin: So cultural evolution can happen a lot faster.

Eliezer: Well, for one thing, I could say—it's an unimpressively nonabstract observation, but this thing (*picks up laptop*) does run at around two billion hertz and this thing (*points at head*) runs at about two hundred hertz.

Robin: Right.

Eliezer: If you can have architectural innovations which merely allow this thing (*picks up laptop*) to do the same sort of thing that this thing (*points to head*) is doing, only a million times faster, then that million times faster means that that thirty-one seconds works out to about a subjective year and all the time between ourselves and Socrates works out to about eight hours. It may look like it's—

Robin: Lots of people have those machines in their basements. You have to imagine that your basement has something better. They have those machines. You have your machines. Your machine has to have this architectural advantage that beats out everybody else's machines in their basements.

Eliezer: Hold on, there's two sort of separate topics here. Previously, you did seem to me to be arguing that we just shouldn't expect that much of a speedup. Then there's the separate question of "Well, suppose the speedup was possible, would one basement get it ahead of other basements?"

Robin: To be clear, the dispute here is—I grant fully that these machines are wonderful and we will move more and more of our powerful content to them and they will execute rapidly and reliably in all sorts of ways to help our economy grow quickly, and in fact, I think it's quite likely that the economic growth rate could accelerate and become much faster. That's with the entire world economy working

together, sharing these things, exchanging them and using them. But now the scenario is, in a world where people are using these as best they can with their best architecture, best software, best approaches for the computers, one guy in a basement has a computer that's not really much better than anybody else's computer in a basement except that it's got this architectural thing that allows it to, within a few weeks, take over the world. That's the scenario.

ELIEZER: Again, you seem to be conceding much more probability. I'm not sure to what degree you think it's likely, but you do seem to be conceding much more probability that there is, in principle, some program where if it was magically transmitted to us, we could take a modern-day large computing cluster and turn it into something that could generate what you call content a million times faster.

To the extent that that is possible, the whole brain-in-a-box scenario thing does seem to become intuitively more credible. To put it another way, if you just couldn't have an architecture better than this (*points to head*), if you couldn't run at faster speeds than this, if all you could do was use the same sort of content that had been laboriously developed over thousands of years of civilization, and there wasn't really any way to generate content faster than that, then the "foom" scenario does go out the window.

If, on the other hand, there's this gap between where we are now and this place where you can generate content millions of times faster, then there is a further issue of whether one basement gets that ahead of other basements, but it suddenly does become a lot more plausible if you had a civilization that was ticking along just fine for thousands of years, generating lots of content, and then something else came

along and just sucked all that content that it was interested in off the Internet, and—

Robin: We've had computers for a few decades now. This idea that once we have computers, innovation will speed up—we've already been able to test that idea, right? Computers are useful in some areas as complementary inputs, but they haven't overwhelmingly changed the growth rate of the economy. We've got these devices. They run a lot faster—where we can use them, we use them—but overall limitations to innovation are much more about having good ideas and trying them out in the right places, and pure computation isn't, in our world, that big an advantage in doing innovation.

Eliezer: Yes, but it hasn't been running this algorithm, only faster. (*Gestures to head.*) It's been running spreadsheet algorithms. I fully agree that spreadsheet algorithms are not as powerful as the human brain. I mean, I don't know if there's any animal that builds spreadsheets, but if they do, they would not have taken over the world thereby.

Robin: Right. When you point to your head, you say, "This algorithm." There's million of algorithms in there. We are slowly making your laptops include more and more kinds of algorithms that are the sorts of things in your head. The question is, will there be some sudden threshold where entire heads go into the laptops all at once, or do laptops slowly accumulate the various kinds of innovations that heads contain?

Eliezer: Let me try to take it down a level in concreteness. The idea is there are key insights. You can use them to build an AI. You've got a "brain in a box in a basement" team. They take the key insights, they build the AI, the AI goes out and sucks a lot of information off

the Internet, duplicating a lot of content that way because it's stored in a form where it can understand it on its own and download it very rapidly and absorb it very rapidly.

Then, in terms of taking over the world, nanotechnological progress is not that far ahead of its current level, but this AI manages to crack the protein folding problem so it can email something off to one of those places that will take an emailed DNA string and FedEx you back the proteins in seventy-two hours. There are places like this. Yes, we have them now.

Robin: So, we grant that if there's a box somewhere that's vastly smarter than anybody on Earth, or vastly smarter than any million people on Earth, then we've got a problem. The question is, how likely is that scenario?

Eliezer: What I'm trying to distinguish here is the question of "Does that potential exist?" versus "Is that potential centralized?" To the extent that that you say, "OK. There would in principle be some way to know enough about intelligence that you could build something that could learn and absorb existing content very quickly."

In other words, I'm trying to separate out the question of "How dumb is this thing (*points to head*); how much smarter can you build an agent; if that agent were teleported into today's world, could it take over?" versus the question of "Who develops it, in what order, and were they all trading insights or was it more like a modern-day financial firm where you don't show your competitors your key insights, and so on, or, for that matter, modern artificial intelligence programs?"

Robin: I grant that a head like yours could be filled with lots more stuff, such that it would be vastly more powerful. I will call most of

that stuff "content," you might call it "architecture," but if it's a million little pieces, architecture is kind of—content. The key idea is, are there one or two things, such that, with just those one or two things, your head is vastly, vastly more powerful?

Eliezer: OK. So what do you think happened between chimps and humans?

Robin: Something happened, something additional. But the question is, how many more things are there like that?

Eliezer: One obvious thing is just the speed. You do—

Robin: Between chimps and humans, we developed the ability to transmit culture, right? That's the obvious explanation for why we've been able to grow faster. Using language, we've been able to transmit insights and accumulate them socially rather than in the genes, right?

Eliezer: Well, people have tried raising chimps in human surroundings, and they absorbed this mysterious capacity for abstraction that sets them apart from other chimps. There's this wonderful book about one of these chimps, Kanzi was his name. Very, very famous chimpanzee, probably the world's most famous chimpanzee, and probably the world's smartest chimpanzee as well. They were trying to teach his mother to do these human things. He was just a little baby chimp, he was watching. He picked stuff up. It's amazing, but nonetheless he did not go on to become the world's leading chimpanzee scientist using his own chimpanzee abilities separately.

If you look at human beings, then we have this enormous processing object containing billions upon billions of neurons, and people still fail the Wason selection task. They cannot figure out which playing card they need to turn over to verify the rule "If a card has an even number on one side, it has a vowel on the other." They can't figure out

which cards they need to turn over to verify whether this rule is true or false.

Robin: Again, we're not distinguishing architecture and content here. I grant that you can imagine boxes the size of your brain that are vastly more powerful than your brain. The question is, what could create a box like that? The issue here is—I'm saying the way something like that happens is through the slow accumulation of improvement over time, the hard way. There's no shortcut of having one magic innovation that jumps you there all at once. I'm saying that—

I wonder if we should ask for questions and see if we've lost the audience by now.

Eliezer: Yeah. It does seem to me that you're sort of equivocating between arguing that the gap doesn't exist or isn't crossable versus saying the gap is crossed in a decentralized fashion. But I agree that taking some sort of question from the audience might help refocus us.

Robin: Help us.

Eliezer: Yes. Does anyone want to . . . ?

Robin: We lost you?

Audience Member: Isn't one of the major advantages . . . ?

Eliezer: Voice, please.

Man 1: Isn't one of the major advantages that humans have over animals the prefrontal cortex? More of the design than the content?

Robin: I don't think we know, exactly.

Woman 1: Robin, you were hypothesizing that it would be a series of many improvements that would lead to this vastly smarter metabrain.

Robin: Right.

Woman 1: But if the idea is that each improvement makes the next improvement that much easier, then wouldn't it quickly, quickly look like just one or two improvements?

Robin: The issue is the spatial scale on which improvement happens. For example, if you look at, say, programming languages, a programming language with a lot of users, compared to a programming language with a small number of users, the one with a lot of users can accumulate improvements more quickly, because there are many . . .

(*Laughter.*)

Robin: There are ways you might resist it too, of course. But there are just many people who could help improve it. Or similarly, with something other that gets used by many users, they can help improve it. It's not just what kind of thing it is, but how large a base of people are helping to improve it.

Eliezer: Robin, I have a slight suspicion that Jane Street Capital is using its own proprietary programming language.

(*Laughter.*)

Robin: Right.

Eliezer: Would I be correct in that suspicion?

Robin: Well, maybe get advantages.

Man 2: It's not proprietary—esoteric.

Robin: Esoteric. But still, it's a tradeoff you have. If you use your own thing, you can be specialized. It can be all yours. But you have fewer people helping to improve it.

If we have the thing in the basement, and it's all by itself, it's not sharing innovations with the rest of the world in some large research community that's building on each other, it's just all by itself, working by itself, it really needs some other advantage that is huge to counter

that. Because otherwise we've got a scenario where people have different basements and different machines, and they each find a little improvement and they share that improvement with other people, and they include that in their machine, and then other people improve theirs, and back and forth, and all the machines get better and faster.

Eliezer: Well, present-day artificial intelligence does not actually look like that. So you think that in fifty years artificial intelligence or creating cognitive machines is going to look very different than it does right now.

Robin: Almost every real industrial process pays attention to integration in ways that researchers off on their own trying to do demos don't. People inventing new cars, they didn't have to make a car that matched a road and a filling station and everything else, they just made a new car and said, "Here's a car. Maybe we should try it." But once you have an automobile industry, you have a whole set of suppliers and manufacturers and filling stations and repair shops and all this that are matched and integrated to each other. In a large, actual economy of smart machines with pieces, they would have standards, and there would be strong economic pressures to match those standards.

Eliezer: Right, so a very definite difference of visualization here is that I expect the dawn of artificial intelligence to look like someone successfully building a first-of-its-kind AI that may use a lot of published insights and perhaps even use some published libraries, but it's nonetheless a prototype, it's a one-of-a-kind thing, it was built by a research project.

And you're visualizing that at the time interesting things start to happen—or maybe even there is no key threshold, because there's no

storm of recursive self-improvements—everyone gets slowly better and better at building smarter and smarter machines. There's no key threshold.

Robin: I mean, it is the sort of Bond villain, Captain Nemo on his own island doing everything, beating out the rest of the world isolated, versus an integrated . . .

Eliezer: Or rise of human intelligence. One species beats out all the other species. We are not restricted to fictional examples.

Robin: Human couldn't share with the other species, so there was a real limit.

Man 3: In one science fiction novel, I don't remember its name, there was a very large swarm of nanobots. These nanobots had been created so long ago that no one knew what the original plans were. You could ask the nanobots for their documentation, but there was no method, they'd sometimes lie. You couldn't really trust the manuals they gave you. I think one question that's happening here is when we have a boundary where we hit the point where suddenly someone's created software that we can't actually understand, like it's not actually within our—

Robin: We're there. (*Laughs.*)

Man 3: Well, so are we actually there . . . So, Hanson—

Robin: We've got lots of software we don't understand. Sure. (*Laughs.*)

Man 3: But we can still understand it at a very local level, we can still disassemble it. It's pretty surprising to what extent Windows has been reverse-engineered by the millions of programmers who work on it. I was going to ask you if getting to that point was key to the resulting exponential growth, which is not permitting the transfer of

information. Because if you can't understand the software, you can't transmit the insights using your own process.

ELIEZER: That's not really a key part of my visualization. I think that there's a sort of mysterian tendency, like people who don't know how neural networks work are very impressed by the fact that you can train neural networks to do something you don't know how it works. As if your ignorance of how they worked was responsible for making them work better somehow. So *ceteris paribus*, not being able to understand your own software is a bad thing.

ROBIN: Agreed.

ELIEZER: I wasn't really visualizing there being a key threshold where incomprehensible software is a . . . Well, OK. The key piece of incomprehensible software in this whole thing is the brain. This thing is not end-user modifiable. If something goes wrong you cannot just swap out one module and plug in another one, and that's why you die. You die, ultimately, because your brain is not end-user modifiable and doesn't have I/O ports or hot-swappable modules or anything like that.

The reason why I expect localist sorts of things is that I expect one project to go over the threshold for intelligence in much the same way that chimps went over the threshold of intelligence and became humans. (Yes, I know that's not evolutionarily accurate.)

Then, even though they now have this functioning mind, to which they can make all sorts of interesting improvements and have it run even better and better . . . Whereas meanwhile all the other cognitive work on the planet is being done by these non-end-user-modifiable human intelligences which cannot really make very good use of the insights, although it is an intriguing fact that after spending some

time trying to figure out artificial intelligence I went off and started blogging about human rationality.

Man 4: I just wanted to clarify one thing. Would you guys both agree—well, I know you would agree—would you agree, Robin, that in your scenario—just imagine one had a time machine that could carry a physical object the size of this room, and you could go forward a thousand years into the future and essentially create and bring back to the present day an object, say, the size of this room, that you could take over the world with that?

Robin: I have no doubt of that.

Man 4: OK. The question is whether that object is—

Eliezer: Point of curiosity. Does this work too? (*Holds up cell phone.*) Object of this size?

Robin: Probably.

Eliezer: Yeah, I figured. (*Laughs.*)

Man 4: The question is, does the development of that object essentially happen in a very asynchronous way, or more broadly?

Robin: I think I should actually admit that there is a concrete scenario that I can imagine that fits much more of his concerns. I think that the most likely way that the content that's in our heads will end up in silicon is something called "whole-brain emulation," where you take actual brains, scan them, and make a computer model of that brain, and then you can start to hack them to take out the inefficiencies and speed them up.

If the time at which it was possible to scan a brain and model it sufficiently was a time when the computer power to actually run those brains was very cheap, then you have more of a computing cost overhang, where the first person who can manage to do that can then make

a lot of them very fast, and then you have more of a risk scenario. It's because, with emulation, there is this sharp threshold. Until you have a functioning emulation, you just have shit, because it doesn't work, and then when you have it work, it works as well as any of you, and you can make lots of it.

Eliezer: Right. So, in other words, we get a centralized economic shock, because there's a curve here that has a little step function in it. If I can step back and describe what you're describing on a higher level of abstraction, you have emulation technology that is being developed all over the world, but there's this very sharp threshold in how well the resulting emulation runs as a function of how good your emulation technology is. The output of the emulation experiences a sharp threshold.

Robin: Exactly.

Eliezer: In particular, you can even imagine there's a lab that builds the world's first correctly functioning scanner. It would be a prototype, one-of-its-kind sort of thing. It would use lots of technology from around the world, and it would be very similar to other technology from around the world, but because they got it, you know, there's one little extra year they added on, they are now capable of absorbing all of the content in here (*points at head*) at an extremely great rate of speed, and that's where the first-mover effect would come from.

Robin: Right. The key point is, for an emulation there's this threshold. If you get it almost right, you just don't have something that works. When you finally get enough, then it works, and you get all the content through. It's like if some aliens were sending a signal and we just couldn't decode their signal. It was just noise, and then finally we figured out the code, and then we've got a high band-

width rate and they're telling us lots of technology secrets. That would be another analogy, a sharp threshold where suddenly you get lots of stuff.

Eliezer: So you think there's a mainline, higher than 50%, probability that we get this sort of threshold with emulations?

Robin: It depends on which is the last technology to be ready with emulations. If computing is cheap when the thing is ready, then we have this risk. I actually think that's relatively unlikely, that the computing will still be expensive when the other things are ready, but . . .

Eliezer: But there'd still be a speed-of-content-absorption effect, it just wouldn't give you lots of emulations very quickly.

Robin: Right. It wouldn't give you this huge economic power.

Eliezer: And similarly, with chimpanzees we also have some indicators that at least their ability to do abstract science . . . There's what I like to call the "one wrong number" function curve or the "one wrong number" curve where dialing 90% of my phone number correctly does not get you 90% of Eliezer Yudkowsky.

Robin: Right.

Eliezer: So similarly, dialing 90% of human correctly does not get you a human—or 90% of a scientist.

Robin: I'm more skeptical that there's this architectural thing between humans and chimps. I think it's more about the social dynamic of "we managed to have a functioning social situation."

Eliezer: Why can't we raise chimps to be scientists?

Robin: Most animals can't be raised to be anything in our society. Most animals aren't domesticatable. It's a matter of whether they evolved the social instincts to work together.

Eliezer: But Robin, do you actually think that if we could domesticate chimps they would make good scientists?

Robin: They would certainly be able to do a lot of things in our society. There are a lot of roles in even scientific labs that don't require that much intelligence.

(*Laughter.*)

Eliezer: OK, so they can be journal editors, but can they actually be innovators? (*Laughs.*)

(*Laughter.*)

Robin: For example.

Man 5: My wife's a journal editor!

(*Laughter.*)

Robin: Let's take more questions.

Eliezer: My sympathies.

(*Laughter.*)

Robin: Questions.

Man 6: Professor Hanson, you seem to have the idea that social skill is one of the main things that separate humans from chimpanzees. Can you envision a scenario where one of the computers acquired this social skill and comes to the other computers and says, "Hey, guys, we can start a revolution here!"?

(*Laughter.*)

Man 6: Maybe that's the first mover, then? That might be the first mover?

Robin: One of the nice things about the vast majority of software in our world is that it's really quite socially compliant. You can take a chimpanzee and bring him in and you can show him some tasks and then he can do it for a couple of hours. Then just some time randomly

in the next week he'll go crazy and smash everything, and that ruins his entire productivity. Software doesn't do that so often.

(*Laughter.*)

Eliezer: No comment. (*Laughs.*)

(*Laughter.*)

Robin: Software, the way it's designed, is set up to be relatively socially compliant. Assuming that we continue having software like that, we're relatively safe. If you go out and design software like wild chimps that can just go crazy and smash stuff once in a while, I don't think I want to buy your software. (*Laughs.*)

Man 7: I don't know if this sidesteps the issue, but to what extent do either of you think something like government classification, or the desire of some more powerful body to innovate and then keep what it innovates secret, could affect centralization to the extent you were talking about?

Eliezer: As far as I can tell, what happens when the government tries to develop AI is nothing, but that could just be an artifact of our local technological level and it might change over the next few decades.

To me it seems like a deeply confusing issue whose answer is probably not very complicated in an absolute sense. We know why it's difficult to build a star. You've got to gather a very large amount of interstellar hydrogen in one place. We understand what sort of labor goes into a star and we know why a star is difficult to build.

When it comes to building a mind, we don't know how to do it, so it seems very hard. We query our brains to say, "Map us a strategy to build this thing," and it returns null, so it feels like it's a very difficult

problem. But in point of fact, we don't actually know that the problem is difficult apart from being confusing.

We understand the star-building problems. We know it's difficult. This one, we don't know how difficult it's going to be after it's no longer confusing. So, to me, the AI problem looks like the problem is finding bright enough researchers, bringing them together, letting them work on that problem instead of demanding that they work on something where they're going to produce a progress report in two years which will validate the person who approved the grant and advance their career.

The government has historically been tremendously bad at producing basic research progress in AI, in part because the most senior people in AI are often people who got to be very senior by having failed to build it for the longest period of time. This is not a universal statement. I've met smart senior people in AI, but nonetheless.

Basically I'm not very afraid of the government because I don't think it's "throw warm bodies at the problem," and I don't think it's "throw warm computers at the problem," I think it's good methodology, good people selection, letting them do sufficiently blue-sky stuff, and so far, historically, the government has just been tremendously bad at producing that kind of progress. When they have a great big project and try to build something, it doesn't work. When they fund long-term research—

Robin: I agree with Eliezer that in general you too often go down the route of trying to grab something before it's grabbable. But there is the scenario—certainly in the midst of a total war, when you have a technology that seems to have strong military applications and not

much other application, you'd be wise to keep that application within the nation or your side of the alliance in the war.

But there's too much of a temptation to use that sort of thinking when you're not in a war, or when the technology isn't directly military-applicable but has several steps of indirection. You can often just screw it up by trying to keep it secret.

That is, your tradeoff is between trying to keep it secret and getting this advantage versus putting this technology into the pool of technologies that the entire world develops together and shares, and usually that's the better way to get advantage out of it unless you can, again, identify a very strong military application with a particular immediate use.

Eliezer: That sounds like a plausible piece of economic logic, but it seems plausible to the same extent as the economic logic which says there should obviously never be wars because they're never Pareto optimal. There's always a situation where you didn't spend any of your resources in attacking each other, which was better. And it sounds like the economic logic which says that there should never be any unemployment because of Ricardo's Law of Comparative Advantage, which means there's always someone who you can trade with.

If you look at the state of present-world technological development, there's basically either published research or proprietary research. We do not see corporations in closed networks where they trade their research with each other but not with the outside world. There's either published research, with all the attendant free-rider problems that implies, or there's proprietary research. As far as I know, may this room correct me if I'm mistaken, there is not a set

of, like, three leading trading firms which are trading all of their internal innovations with each other and not with the outside world.

Robin: If you're a software company, and you locate in Silicon Valley, you've basically agreed that a lot of your secrets will leak out as your employees come in and leave your company. Choosing where to locate a company is often a choice to accept a certain level of leakage of what happens within your company in trade for a leakage from the other companies back toward you. So, in fact, people who choose to move to those areas in those industries do in fact choose to have a set of . . .

Eliezer: But that's not trading innovations with each other and not with the rest of the outside world. I can't actually even think of where we would see that pattern.

Robin: It is. More trading with the people in the area than with the rest of the world.

Eliezer: But that's coincidental side-effect trading. That's not deliberate, like, "You scratch my back . . ."

Robin: But that's why places like that get the big advantage, because you go there and lots of stuff gets traded back and forth.

Eliezer: Yes, but that's the commons. It's like a lesser form of publication. It's not a question of me offering this company an innovation in exchange for their innovation.

Robin: Well, we're probably a little sidetracked. Other . . .

Man 8: It's actually relevant to this little interchange. It seems to me that there's both an economic and social incentive for people to release partial results and imperfect products and steps along the way, which it seems would tend to yield a more gradual approach towards

this breakthrough that we've been discussing. Do you disagree? I know you disagree, but why do you disagree?

Eliezer: Well, here at the Singularity Institute, we plan to keep all of our most important insights private and hope that everyone else releases their results.

(*Laughter.*)

Man 8: Right, but most human-inspired innovations haven't worked that way, which then I guess—

Eliezer: Well, we certainly hope everyone else thinks that way.

(*Laughter.*)

Robin: Usually you'll have a policy about having these things leaked, but in fact you make very social choices that you know will lead to leaks, and you accept those leaks in trade for the other advantages those policies bring. Often they are that you are getting leaks from others. So locating yourself in a city where there are lots of other firms, sending your people to conferences where other people going to the same conferences, those are often ways in which you end up leaking and getting leaks in trade.

Man 8: So the team in the basement won't release anything until they've got the thing that's going to take over the world?

Eliezer: Right. We were not planning to have any windows in the basement.

(*Laughter.*)

Man 9: Why do we think that . . .

Eliezer: If anyone has a microphone that can be set up over here, I will happily donate this microphone.

Man 9: Why do we think that, if we manage to create an artificial human brain, that it would immediately work much, much faster than

a human brain? What if a team in the basement makes an artificial human brain, but it works at one billionth the speed of a human brain? Wouldn't that give other teams enough time to catch up?

Eliezer: First of all, the course we're visualizing is not like building a human brain in your basement, because, based on what we already understand about intelligence . . . We don't understand everything, but we understand some things, and what we understand seems to me to be quite sufficient to tell you that the human brain is a completely crap design, which is why it can't solve the Wason selection task.

You pick up any bit of the heuristics and biases literature and there's one hundred different ways that this thing reliably experimentally malfunctions when you give it some simple-seeming problems. You wouldn't want to actually want to build anything that worked like the human brain. It would miss the entire point of trying to build a better intelligence.

But if you were to scan a brain—this is something that Robin has studied in more detail than I have—then the first one might run at one thousandth your speed or it might run at one thousand times your speed. It depends on the hardware overhang, on what the cost of computer power happens to be at the point where your scanners get good enough. Is that fair?

Robin: Or your modeler is good enough.

Actually, the scanner being the last thing isn't such a threatening scenario because then you'd have a big consortium get together to do the last scan when it's finally cheap enough. But the modeling being the last thing is more disruptive, because it's just more uncertain when modeling gets done.

Eliezer: By modeling you mean?

Robin: The actual modeling of the brain cells in terms of translating a scan into—

Eliezer: Oh, I see. So in other words, if there's known scans but you can't model the brain cells, then there's an even worse last-mile problem?

Robin: Exactly.

Eliezer: I'm trying to think if there's anything else I can . . .

I would hope to build an AI that was sufficiently unlike human, because it worked better, that there would be no direct concept of "How fast does this run relative to you?" It would be able to solve some problems very quickly, and if it can solve all problems much faster than you, we're already getting into the superintelligence range.

But at the beginning, you would already expect it to be able to do arithmetic immensely faster than you, and at the same time it might be doing basic scientific research a bit slower. Then eventually it's faster than you at everything, but possibly not the first time you boot up the code.

Man 10: I'm trying to envision intelligence explosions that win Robin over to Yudkowsky's position. Does either one of these, or maybe a combination of both, self-improving software or nanobots that build better nanobots, is that unstable enough? Or do you still sort of feel that would be a widespread benefit?

Robin: The key debate we're having isn't about the rate of change that might eventually happen. It's about how local that rate of change might start.

If you take the self-improving software—of course, we have software that self-improves, it just does a lousy job of it. If you imagine

steady improvement in the self-improvement, that doesn't give a local team a strong advantage. You have to imagine that there's some clever insight that gives a local team a vast, cosmically vast, advantage in its ability to self-improve compared to the other teams such that not only can it self-improve, but it self-improves like gangbusters in a very short time.

With nanobots again, if there's a threshold where you have nothing like a nanobot and then you have lots of them and they're cheap, that's more of a threshold kind of situation. Again, that's something that the nanotechnology literature had a speculation about a while ago. I think the consensus moved a little more against that in the sense that people realized those imagined nanobots just wouldn't be as economically viable as some larger-scale manufacturing process to make them.

But again, it's the issue of whether there's that sharp threshold where you're almost there and it's just not good enough because you don't really have anything, and then you finally pass the threshold and now you've got vast power.

Eliezer: What do you think you know and how do you think you know it with respect to this particular issue of [whether] that which yields the power of human intelligence is made up of a thousand pieces, or a thousand different required insights? Is this something that should seem more plausible in principle? Where does that actually come from?

Robin: One set of sources is just what we've learned as economists and social scientists about innovation in our society and where it comes from. That innovation in our society comes from lots of little things accumulating together, it rarely comes from one big thing.

It's usually a few good ideas and then lots and lots of detail worked out. That's generically how innovation works in our society and has for a long time. That's a clue about the nature of what makes things work well, that they usually have some architecture and then there's just lots of detail and you have to get it right before something really works.

Then, in the AI field in particular, there's also this large . . . I was an artificial intelligence researcher for nine years, but it was a while ago. In that field in particular there's this . . . The old folks in the field tend to have a sense that people come up with new models. But if you look at their new models, people remember a while back when people had something a lot like that, except they called it a different name. And they say, "Fine, you have a new name for it."

You can keep reinventing new names and new architectures, but they keep cycling among a similar set of concepts for architecture. They don't really come up with something very dramatically different. They just come up with different ways of repackaging different pieces in the architecture for artificial intelligence. So there was a sense to which—maybe we'll find the right combination but it's clear that there's just a lot of pieces together.

In particular, Douglas Lenat did this system that you and I both respect called Eurisko a while ago that had this nice simple architecture and was able to self-modify and was able to grow itself, but its growth ran out and slowed down. It just couldn't improve itself very far even though it seemed to have a nice, elegant architecture for doing so. Lenat concluded, and I agree with him, that the reason it couldn't go very far is it just didn't know very much. The key to making something like that work was to just collect a lot more knowledge

and put it in so it had more to work with when it was trying to modify and make improvements.

ELIEZER: But Lenat's still trying to do that fifteen years later and so far Cyc does not seem to work even as well as EURISKO.

ROBIN: Cyc does some pretty impressive stuff. I'll agree that it's not going to replace humans any time soon, but it's an impressive system, if you look at it.

ELIEZER: It seems to me that Cyc is an iota of evidence against this view. That's what Cyc was supposed to do. You're supposed to put in lots of knowledge and then it was supposed to go foom, and it totally didn't.

ROBIN: It was supposed to be enough knowledge and it was never clear how much is required. So apparently what they have now isn't enough.

ELIEZER: But clearly Lenat thought there was some possibility it was going to go foom in the next fifteen years. It's not that this is quite unfalsifiable, it's just been incrementally more and more falsified.

ROBIN: I can point to a number of senior AI researchers who basically agree with my point of view that this AI foom scenario is very unlikely. This is actually more of a consensus, really, among senior AI researchers.

ELIEZER: I'd like to see that poll, actually, because I could point to AI researchers who agree with the opposing view as well.

ROBIN: AAAI has a panel where they have a white paper where they're coming out and saying explicitly, "This explosive AI view, we don't find that plausible."

ELIEZER: Are we talking about the one with, what's his name, from . . . ?

Robin: Norvig?

Eliezer: Eric Horvitz?

Robin: Horvitz, yeah.

Eliezer: Was Norvig on that? I don't think Norvig was on that.

Robin: Anyway, Norvig just made the press in the last day or so arguing about linguistics with Chomsky, saying that this idea that there's a simple elegant theory of linguistics is just wrong. It's just a lot of messy detail to get linguistics right, which is a similar sort of idea. There is no key architecture—

Eliezer: I think we have a refocusing question from the audience.

Man 11: No matter how smart this intelligence gets, to actually take over the world . . .

Eliezer: Wait for the microphone. Wait for the microphone.

Man 11: This intelligence has to interact with the world to be able to take it over. So if we had this box, and we were going to use it to try to make all the money in the world, we would still have to talk to all the exchanges in the world, and learn all the bugs in their protocol, and the way that we're able to do that is that there are humans at the exchanges that operate at our frequency and our level of intelligence, we can call them and ask questions.

And this box, if it's a million times smarter than the exchanges, it still has to move at the speed of the exchanges to be able to work with them and eventually make all the money available on them. And then if it wants to take over the world through war, it has to be able to build weapons, which means mining, and building factories, and doing all these things that are really slow and also require extremely high-dimensional knowledge that seems to have nothing to do with

just how fast you can think. No matter how fast you can think, it's going to take a long time to build a factory that can build tanks.

How is this thing going to take over the world when . . . ?

Eliezer: The analogy that I use here is, imagine you have two people having an argument just after the dawn of human intelligence. There's these two aliens in a spaceship, neither of whom have ever seen a biological intelligence—we're going to totally skip over how this could possibly happen coherently. But there are these two observers in spaceships who have only ever seen Earth, and they're watching these new creatures who have intelligence. They're arguing over, how fast can these creatures progress?

One of them says, "Well, it doesn't matter how smart they are. They've got no access to ribosomes. There's no access from the brain to the ribosomes. They're not going to be able to develop new limbs or make honey or spit venom, so really we've just got these squishy things running around without very much of an advantage for all their intelligence, because they can't actually make anything, because they don't have ribosomes."

And we eventually bypassed that whole sort of existing infrastructure and built our own factory systems that had a more convenient access to us. Similarly, there's all this sort of infrastructure out there, but it's all infrastructure that we created. The new system does not necessarily have to use our infrastructure if it can build its own infrastructure.

As for how fast that might happen, well, in point of fact we actually popped up with all these factories on a very rapid timescale compared to the amount of time it took natural selection to produce ribosomes.

We were able to build our own new infrastructure much more quickly than it took to create the previous infrastructure.

To put it on a very concrete level, if you can crack the protein folding problem, you can email a DNA string to one of these services that will send you back the proteins that you asked for with a seventy-two-hour turnaround time. Three days may sound like a very short period of time to build your own economic infrastructure relative to how long we're used to it taking, but in point in fact this is just the cleverest way that I could think of to do it, and seventy-two hours would work out to I don't even know how long at a million-to-one speedup rate. It would be like thousands upon thousands upon thousands of years. But there might be some even faster way to get your own infrastructure than the DNA . . .

Man 11: Is this basic argument something you two roughly agree on or roughly disagree on?

Robin: I think we agree on the specific answer to the question, but we differ on how to frame it. I think it's relevant to our discussion. I would say our civilization has vast capacity and most of the power of that capacity is a mental capacity. We, as a civilization, have a vast mental capacity. We are able to think about a lot of things and calculate and figure out a lot of things.

If there's a box somewhere that has a mental capacity comparable to the rest of human civilization, I've got to give it some respect and figure it can do a hell of a lot of stuff. I might quibble with the idea that if it were just intelligent it would have that mental capacity. Because it comes down to, well, this thing was improving what about itself exactly? So there's the issue of, what various kinds of things does it take to produce various kinds of mental capacities?

I'm less enamored of the idea that there's this intelligence thing. If it's just intelligent enough it doesn't matter what it knows, it's just really smart. And I'm not sure that concept makes sense. I'm happy to grant the idea that if—

ELIEZER: Or it can learn much faster than you can learn. It doesn't necessarily have to go through college the way you did, because it is able to, much more rapidly, learn either by observing reality directly or, in point of fact, given our current state of society, you can just cheat, you can just download it from the Internet.

ROBIN: Simply positing it has a great mental capacity, then I will be in fear of what it does. The question is, how does it get that capacity?

ELIEZER: Would the audience be terribly offended if I tried to answer that one a bit? The thing is there is a number of places the step function can come in. We could have a historical step function like what happens from humans to chimps. We could have the combined effect of all the obvious ways to rebuild an intelligence if you're not doing it evolutionarily.

You build an AI and it's on a 2 GHz chip instead of 200 Hz neurons. It has complete read and write access to all the pieces of itself. It can do repeatable mental processes and run its own internal, controlled experiments on what sort of mental processes work better and then copy it onto new pieces of code. Unlike this hardware (*points to head*) where we're stuck with a certain amount of hardware, if this intelligence works well enough it can buy, or perhaps simply steal, very large amounts of computing power from the large computing clusters that we have out there.

If you want to solve a problem, there's no way that you can allocate, reshuffle, reallocate internal resources to different aspects of it. To me it looks like, architecturally, if we've got down the basic insights that underlie human intelligence, and we can add all the cool stuff that we could do if we were designing an artificial intelligence instead of being stuck with the ones that evolution accidentally burped out, it looks like they should have these enormous advantages.

We may have six billion people on this planet, but they don't really add that way. Six billion humans are not six billion times as smart as one human. I can't even imagine what that planet would look like. It's been known for a long time that buying twice as many researchers does not get you twice as much science. It gets you twice as many science papers. It does not get you twice as much scientific progress.

Here we have some other people in the Singularity Institute who have developed theses that I wouldn't know how to defend myself, which are more extreme than mine, to the effect that if you buy twice as much science you get flat output or even it actually goes down because you increase the signal-to-noise ratio. But now I'm getting a bit off track.

Where does this enormous power come from? It seems like human brains are just not all that impressive. We don't add that well. We can't communicate with other people. One billion squirrels could not compete with the human brain. Our brain is about four times as large as a chimp's, but four chimps cannot compete with one human.

Making a brain twice as large and actually incorporating it into the architecture seems to produce a scaling of output of intelligence that is not even remotely comparable to the effect of taking two brains of fixed size and letting them talk to each other using words. So an

artificial intelligence that can do all this neat stuff internally and possibly scale its processing power by orders of magnitude, that itself has a completely different output function than human brains trying to talk to each other.

To me, the notion that you can have something incredibly powerful and, yes, more powerful than our sad little civilization of six billion people flapping their lips at each other running on 200 Hz brains, is actually not all that implausible.

Robin: There are devices that think, and they are very useful. So 70% of world income goes to pay for creatures who have these devices that think, and they are very, very useful. It's more of an open question, though, how much of that use is because they are a generic good thinker or because they know many useful particular things?

I'm less assured of this idea that you just have a generically smart thing and it's not smart about anything at all in particular. It's just smart in the abstract. And that it's vastly more powerful because it's smart in the abstract compared to things that know a lot of concrete things about particular things.

Most of the employees you have in this firm or in other firms, they are useful not just because they were generically smart creatures but because they learned a particular job. They learned about how to do the job from the experience of other people, on the job, and practice, and things like that.

Eliezer: Well, no. First you needed some very smart people and then you taught them the job. I don't know what your function over here looks like, but I suspect if you take a bunch of people who are thirty IQ points down the curve and try to teach them the same job—I'm not quite sure what would happen then, but I would guess that

your corporation would probably fall a bit in the rankings of financial firms, however those get computed.

Robin: So there's the question of what it means to be smarter.

Eliezer: And thirty IQ points is just like this tiny little mental difference compared to any of the actual "we are going to reach in and change around the machinery and give you different brain areas." Thirty IQ points is nothing and yet it seems to make this very large difference in practical output.

Robin: When we look at people's mental abilities across a wide range of tasks, we do a factor analysis of that, we get the dominant factor, the eigenvector with the biggest eigenvalue, and that we call intelligence. It's the one-dimensional thing that explains the most correlation across different tasks. It doesn't mean that there is therefore an abstract thing that you can build into an abstract thing, a machine, that gives you that factor. It means that actual real humans are correlated in that way. And then the question is, what causes that correlation?

There are many plausible things. One, for example, is simply assortative mating. People who are smart in some ways mate with other people smart in other ways, that produces a correlation across . . . Another could be there's just an overall strategy that some minds devote more resources to different kinds of tasks. There doesn't need to be any central abstract thing that you can make a mind do that lets it solve lots of problems simultaneously for there to be this IQ factor of correlation.

Eliezer: So then why humans? Why weren't there twenty different species that got good at doing different things?

Robin: We grant that there is something that changed with humans, but that doesn't mean that there's this vast landscape of intelligence you can create that's billions of times smarter than us just by rearranging the architecture. That's the key thing.

Eliezer: It seems to me that for this particular argument to carry, it's not enough to say you need content. There has to be no master trick to learning or producing content. And there in particular I can't actually say, "Bayesian updating," because doing it on the full distribution is not computationally tractable. You need to be able to approximate it somehow.

Robin: Right.

Eliezer: But nonetheless there's this sort of core trick called learning, or Bayesian updating. And you look at human civilization and there's this core trick called science. It's not that the science of figuring out chemistry was developed in one place and it used something other than the experimental method compared to the science of biology that was developed in another place. Sure, there were specialized skills that were developed afterward. There was also a core insight, and then people practiced the core insight and they started developing further specialized skills over a very short timescale compared to previous civilizations before that insight had occurred.

It's difficult to look over history and think of a good case where there has been . . . Where is the absence of the master trick which lets you rapidly generate content? Maybe the agricultural revolution. Maybe for the agricultural revolution . . . Well, even for the agricultural revolution, first there's the master trick, "I'm going to grow plants," and then there's developing skills at growing a bunch of different plants.

Robin: There's a large literature on technological and economic innovation, and it basically says the vast majority of innovation is lots of small gains. You can look at locomotives and when locomotives got faster and more energy efficient. You could look at lots of particular devices, and basically you do some curve of how well they got over time, and it's basically lots of little steps over time that slowly made them better.

Eliezer: Right. But this is what I expect a superintelligence to look like after the sort of initial self-improvement passes and it's doing incremental gains. But in the beginning, there's also these very large insights.

Robin: That's what we're debating. Other questions or comments?

Moderator: Actually, before—Craig, you can take this—can everybody without making a big disruption pass your votes to this side of the room and we can tabulate them and see what the answers are. But continue with the questions.

Eliezer: Remember, "yes" is this side of the room and "no" is that side of the room.

(*Laughter.*)

Man 12: I just wanted to make sure I understood the relevance of some of the things we're talking about. I think you both agree that if the time it takes to get from a machine that's, let's say, a tenth as effective as humans to, let's say, ten times as effective as humans at whatever these being-smart tasks are, like making better AI or whatever—that if that time is shorter, then it's more likely to be localized? Just kind of the sign of the derivative there, is that agreed upon?

Eliezer: I think I agree with that.

Man 12: You agree with it.

Robin: I think when you hypothesize this path of going from one-tenth to ten times—

Eliezer: Robin, step up to the microphone.

Robin:—are you hypothesizing a local path where it's doing its own self-improvement, or are you hypothesizing a global path where all machines in the world are getting better?

Man 12: Let's say that . . .

Eliezer: Robin, step towards the microphone.

Robin: Sorry. (*Laughs.*)

Man 12: Let's say it just turns out to take a fairly small amount of time to get from that one point to the other point.

Robin: But it's a global process?

Man 12: No, I'm saying, how does the fact that it's a short amount of time affect the probability that it's local versus global? Like if you just received that knowledge.

Robin: On time it would be the relative scale of different timescales. If it takes a year but we're in a world economy that doubles every month, then a year is a long time. You have to compare that timescale—

Man 12: I'm talking about from one-tenth human power to ten times. I think we're not yet . . . we probably don't have an economy at that point that's doubling every month, at least not because of AI.

Robin: The point is, if that's a global timescale, if the world is . . . if new issues are showing up every day that are one percent better, then that adds up to that over a period of a year. But everybody shares those innovations every day, then we have a global development. If we've got one group that has a development and jumps a factor of two

all by itself without any other inputs, then you've got a more local development.

Eliezer: Is there any industry in which there's a group of people who share innovations with each other and who could punish someone who defected by using the innovations without publishing their own? Is there any industry that works like that?

Robin: But in all industries, in fact, there's a lot of leakage. This is just generically how industries work, how innovation works in our world. People try to keep things secret, but they fail and things leak out. So teams don't, in fact, get that much further ahead of other teams.

Eliezer: But if you're willing to spend a bit more money you can keep secrets.

Robin: Why don't they, then? Why don't firms actually keep more secrets?

Eliezer: The NSA actually does, and they succeed.

Man 12: So in summary, you thought it was more likely to be local if it happens faster. You didn't think the opposite—

Robin: It depends on what else you're holding constant. Obviously I agree that, holding all the other speeds constant, making that faster makes it more likely to be local.

Eliezer: OK, so holding all other speeds constant, increasing the relative speed of something makes it more likely to be local.

Robin: Right.

Man 12: OK. And that's where we get the relevance of whether it's one or two or three key insights versus if it's lots of small things? Because lots of small things will take more time to accumulate.

Robin: Right. And they leak.

Man 12: So in some sense it's easier to leak one key idea like—

Robin: But when?

Man 12:—like Gaussian processes or something, than it is to leak—

Eliezer: Shh!

Man 12: a vast database of . . .

(*Laughter.*)

Man 12: . . . knowledge that's all kind of linked together in a useful way.

Robin: Well, it's not about the timescale of the leak. So you have some insights, you have thirty of them that other people don't have, but they have thirty that you don't, so you're leaking and they're spreading across. Your sort of overall advantage might be relatively small, even though you've got thirty things they don't. There's just lots of different ones. When there's one thing, and it's the only one thing that matters, then it's more likely that one team has it and other ones don't at some point.

Eliezer: Maybe the singulars who will have five insights, and then the other ten insights or whatever, would be published by industry or something? By people who didn't quite realize that who has these insights is an issue? I mean, I would prefer more secrecy generally, because that gives more of an advantage to localized concentrations of intelligence, which makes me feel slightly better about the outcome.

Robin: The main issue here clearly has to be, how different is this technology from other ones? If we are willing to posit that this is like other familiar technologies, we have a vast experience based on how often one team gets how far ahead of another.

Eliezer: And they often get pretty darn far. It seems to me like the history of technology is full of cases where one team gets way, way, way ahead of another team.

Robin: Way ahead on a relatively narrow thing. You're imagining getting way ahead on the entire idea of mental capacity.

Eliezer: No, I'm just imagining getting ahead on––

Robin: Your machine in the basement gets ahead on everything.

Eliezer: No, I'm imagining getting ahead on this relatively narrow, single technology of intelligence. (*Laughs.*)

Robin: I think intelligence is like "betterness," right? It's a name for this vast range of things we all care about.

Eliezer: And I think it's this sort of machine which has a certain design and churns out better and better stuff.

Robin: But there's this one feature called "intelligence."

Eliezer: Well, no. It's this machine you build. Intelligence is described through work that it does, but it's still like an automobile. You could say, "What is this mysterious forwardness that an automobile possesses?"

Robin: New York City is a good city. It's a great city. It's a better city. Where do you go to look to see the betterness of New York City? It's just in thousands of little things. There is no one thing that makes New York City better.

Eliezer: Right. Whereas I think intelligence is more like a car, it's like a machine, it has a function, it outputs stuff. It's not like a city that's all over the place.

(*Laughter.*)

Man 13: If you could take a standard brain and run it twenty times faster, do you think that's probable? Do you think that won't

happen in one place suddenly? If you think that it's possible, why don't you think it'll lead to a local "foom"?

Robin: So now we're talking about whole-brain emulation scenarios? We're talking about brain scans, then, right?

Man 13: Sure. Just as a path to AI.

Robin: If artificial emulations of brains can run twenty times faster than human brains, but no one team can make their emulations run twenty times more cost-effectively than any of the other teams' emulations, then you have a new economy with cheaper emulations, which is more productive, grows faster, and everything, but there's not a local advantage that one group gets over another.

Eliezer: I don't know if Carl Shulman talked to you about this, but I think he did an analysis suggesting that, if you can run your ems 10% faster, then everyone buys their ems from you as opposed to anyone else. Which is itself contradicted to some extent by a recent study, I think it was a McKinsey study, showing that productivity varies between factories by a factor of five and it still takes ten years for the less efficient ones to go out of business.

Robin: That was on my blog a few days ago.

Eliezer: Ah. That explains where I heard about it. (*Laughs.*)

Robin: Of course.

Eliezer: But nonetheless, in Carl Shulman's version of this, whoever has ems 10% faster soon controls the entire market. Would you agree or disagree that that is likely to happen?

Robin: I think there's always these fears that people have that if one team we're competing with gets a little bit better on something, then they'll take over everything. But it's just a lot harder to take over everything because there's always a lot of different dimensions

on which things can be better, and it's hard to be consistently better in a lot of things all at once. Being 10% better at one thing is not usually a huge advantage. Even being twice as good at one thing is not often that big an advantage.

Eliezer: And I think I'll actually concede the point in real life, but only because the market is inefficient.

Robin: Behind you.

Moderator: We're . . .

Robin: Out of time?

Moderator: Yeah. I think we try to keep it to ninety minutes and you both have done a great job. Maybe take a couple minutes each to—

Robin: What's the vote?

Moderator: I have the results. The pre-wrapping-up comments, but do you both want to maybe three minutes to sum up your view, or do you just want to pull the plug?

Robin: Sure.

Eliezer: Sure.

Robin: I respect Eliezer greatly. He's a smart guy. I'm glad that, if somebody's going to work on this problem, it's him. I agree that there is a chance that it's real. I agree that somebody should be working on it. The issue on which we disagree is, how large a probability is this scenario relative to other scenarios that I fear get neglected because this one looks so sexy?

There is a temptation in science fiction and in lots of fiction to imagine that this one evil genius in the basement lab comes up with this great innovation that lets them perhaps take over the world unless

Bond sneaks in and listens to his long speech about why he's going to kill him, *et cetera*.

(*Laughter.*)

It's just such an attractive fantasy, but that's just not how innovation typically happens in the world. Real innovation has lots of different sources, usually lots of small pieces. It's rarely big chunks that give huge advantages.

Eventually we will have machines that will have lots of mental capacity. They'll be able to do a lot of things. We will move a lot of the content we have in our heads over to these machines. But I don't see the scenario being very likely whereby one guy in a basement suddenly has some grand formula, some grand theory of architecture, that allows this machine to grow from being a tiny thing that hardly knows anything to taking over the world in a couple weeks. That requires such vast, powerful architectural advantages for this thing to have that I just don't find it very plausible. I think it's possible, just not very likely. That's the point on which, I guess, we disagree.

I think more attention should go to other disruptive scenarios, whether they're emulations—maybe there'd be a hardware overhang—and other big issues that we should take seriously in these various disruptive future scenarios. I agree that growth could happen very quickly. Growth could go more quickly on a world scale. The issue is, how local will it be?

Eliezer: It seems to me that this is all strongly dependent, first, on the belief that the causes of intelligence get divided up very finely into lots of little pieces that get developed in a wide variety of different places, so that nobody gets an advantage. And second, that if you do get a small advantage, you're only doing a very small fraction of

the total intellectual labor going to the problem. So you don't have a nuclear-pile-gone-critical effect, because any given pile is still a very small fraction of all the thinking that's going into AI everywhere.

I'm not quite sure to say besides, when I look at the world, it doesn't actually look like the world looks like that. I mean, there aren't twenty different species, all of whom are good at different aspects of intelligence and have different advantages. The *g* factor's pretty weak evidence, but it exists. The people talking about *g* factor do seem to be winning on the experimental predictions test versus the people who previously went around talking about multiple intelligences.

It's not a very transferable argument, but to the extent that I actually have a grasp of cognitive science and can try to figure out how this works, it does not look like it's sliced into lots of little pieces. It looks like there's a bunch of major systems doing particular tasks, and they're all cooperating with each other. It's sort of like we have *a* heart, and not one hundred little mini-hearts distributed around the body. It might have been a sort of better system, but nonetheless we just have one big heart over there.

It looks to me like human intelligence is like . . . that there's really obvious, hugely important things you could do with the first prototype intelligence that actually worked. I expect that the critical thing is going to be the first prototype intelligence that actually works and runs on a 2 GHz processor, and can do little experiments to find out which of its own mental processes work better, and things like that.

The first AI that really works is already going to have a pretty large advantage relative to the biological system, so the key driver of change looks more like somebody builds a prototype, and not like this large existing industry reaches a certain quality level at the point where it

is being mainly driven by incremental improvements leaking out of particular organizations.

There are various issues we did not get into at all, like the extent to which this might still look like a bad thing or not from a human perspective, because even if it's nonlocal, there's still this particular group that got left behind by the whole thing, which was the ones with the biological brains that couldn't be upgraded at all. (*Points at head.*) And various other things, but I guess that's mostly my summary of where this particular debate seems to stand.

Robin: Honored to debate you.

(*Applause.*)

Eliezer: Thank you very much.

Robin: And the winner is . . . ?

Moderator: OK so, in this highly unscientific tally with a number of problems, we started off with forty-five for and forty against. I guess unsurprisingly, very compelling arguments from both parts, fewer people had an opinion.

(*Laughter.*)

Moderator: So now we've gone to thirty-three against and thirty-two for, so "against" lost seven and "for" lost thirteen. We have a lot more undecided people than before—

Robin: Good. You should be undecided.

Moderator: —so "against" has it. Thank you very much.

Robin: You're welcome.

(*Applause.*)

59

Debating Yudkowsky

Robin Hanson

3 July 2011

On Wednesday I debated my ex-co-blogger Eliezer Yudkowsky at a private Jane Street Capital event (crude audio here, from 4:45; better video here, transcript here).

I "won" in the sense of gaining more audience votes—the vote was 45–40 (him to me) before, and 32–33 after the debate. That makes me two for two, after my similar "win" over Bryan Caplan (42–10 before, 25–20 after). This probably says little about me, however, since contrarians usually "win" such debates.

Our topic was: *Compared to the farming and industrial revolutions, intelligence-explosion first movers will quickly control a much larger fraction of their new world*. He was pro, I was con. We also debated this subject here on *Overcoming Bias* from June to December 2008. Let me now try to summarize my current position.

The key issue is: how chunky and powerful are as-yet-undiscovered insights into the architecture of "thinking" in general (vs. on particular topics)? Assume there are many such insights, each requiring that brains be restructured to take advantage. (Ordinary humans couldn't use them.) Also assume that the field of AI research reaches a key pivotal level of development. And at that point, imagine some AI research team discovers a powerful insight and builds an AI with an architecture embodying it. Such an AI might then search for more such insights more efficiently than all other the AI research teams who share their results put together.

This new fast AI might then use its advantage to find another powerful insight, restructure itself to take advantage of it, and so on until it was fantastically good at thinking in general. (Or if the first insight were superpowerful, it might jump to this level in one step.) How good? So good that it could greatly outcompete the *entire rest of the world* at the key task of learning the vast ocean of specific knowledge and insights useful for functioning in the world. So good that even though it started out knowing almost nothing, after a few weeks it knows more than the entire rest of the world put together.

(Note that the advantages of silicon and self-modifiable code over biological brains do not count as relevant chunky architectural insights—they are available to all competing AI teams.)

In the debate, Eliezer gave six reasons to think very powerful brain architectural insights remain undiscovered:

1. Human mind abilities have a strong common IQ factor.

2. Humans show many specific mental failings in reasoning.

3. Humans have completely dominated their chimp siblings.
4. Chimps can't function as "scientists" in human society.
5. *Science* was invented, allowing progress in diverse fields.
6. AGI researchers focus on architectures, share little content.

My responses:

1. Human mental abilities correlate across diverse tasks, but this can result from assortative mating (Wikipedia), from task ability complementarities, or from an overall brain chemistry resource parameter. There is little reason to believe high IQ folks have a brain architecture feature that low IQ folks lack.
2. Mind design must trade reliability and accuracy for speed and cost. It is not clear that humans suffer greatly in typical real choices from their many biases. Yes, future brains with lower compute costs will have higher reliability. But this is hardly a new architectural insight.
3. The key human advantage was accumulating insights via culture. Yes, chimps have "culture," but not enough. Humans had more precise and portable culture via language, and more use for it due to free hands and wider ranges. Culture has a threshold effect of giving only minor benefits until it has *enough* support. And in contrast to the farming and industrial revolutions, where second movers still made big gains, chimps couldn't copy or complement humans enough to gain from humans getting culture first. No big architectural advantages are needed to explain human domination.

4. Low-IQ humans also can't function at top levels of human society, and we have no reason to believe they lack some special architecture that the high-IQ have. Chimps' inability to function at our society's low levels, where their intelligence seems plenty sufficient, is explained by only a tiny fraction of animal species ever being domesticated. Most animals refuse to take our orders, even when they are plenty smart enough to understand them.

5. The intellectual community called "science" required a sufficient scale of people, communication, and activity to be feasible. Similar behavior was probably tried many times before, but at insufficient scale. Science required no brain architecture changes.

6. The vast majority of AI researchers focus on collecting and implementing small insights. The fact that a small community of AGI (Artificial General Intelligence) researchers focus on architecture hardly says architecture gives huge gains. And academia discourages the large team projects needed to integrate a lot of content—it is hard to publish on small local changes to large projects.

My five reasons to think powerful architectural insights are quite rare:

1. The literature on economic, technical, and other innovation says most value comes from many small innovations—more useful and wider-scope innovations are rarer, and usually require many small supporting innovations. "Intelligence" covers an *extremely* wide scope, basically all mental tasks. In gen-

eral, innovations come from diverse users and builders, so the more users the better.

2. Whatever appeared first in humans gave them no immediate gains in their ability to support a larger population, but only increased the growth rate of that ability. The same held in the farming and industrial revolutions, the two other most disruptive events by far in human history. The key to all these changes seems to be better ways to spread innovations further faster. Thus any brain architectural gains must have focused mainly on spreading innovations.

3. The usual lore among older artificial intelligence researchers is that new proposed architectural concepts are almost always some sort of rearranging of older architectural concepts. They see little new under the AI sun.

4. The AI system Eliezer most respects for its promising architecture is EURISKO. Its author, Doug Lenat, concluded from it that our main obstacle is not architecture but mental content—the more one knows, the faster one can learn. Lenat's new Cyc system has much content, though it still doesn't learn fast. Cyc might not have enough content yet, or perhaps Lenat sought the wrong content or format.

5. Most AI successes come when hardware costs fall enough to implement old methods more vigorously. Most recent big AI successes are due to better ability to integrate a diversity of small contributions. See how Watson won,[1] or Peter Norvig on

> mass data beating elegant theories.[2] New architecture deserves only small credit for recent success.

Future superintelligences will exist, but their vast and broad mental capacities will come mainly from vast mental content and computational resources. By comparison, their general architectural innovations will be minor additions. It thus seems quite unlikely that one AI team could find an architectural innovation powerful enough to let it go from tiny to taking over the world within a few weeks.

* * *

See original post for all comments.

* * *

1. John Markoff, "Computer Wins on 'Jeopardy!': Trivial, It's Not," *New York Times*, February 16, 2011, http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html.

2. Peter Norvig, "On Chomsky and the Two Cultures of Statistical Learning," May 27, 2011, accessed July 28, 2013, http://norvig.com/chomsky.html.

60

Foom Debate, Again

Robin Hanson

18 February 2013

My ex-co-blogger Eliezer Yudkowsky last June:

> I worry about conversations that go into "But X is like Y, which does Z, so X should do reinterpreted-Z." Usually, in my experience, that goes into what I call "reference class tennis" or "I'm taking my reference class and going home." The trouble is that there's an unlimited number of possible analogies and reference classes, and everyone has a different one. I was just browsing old *LW* posts today (to find a URL of a quick summary of why group-selection arguments don't work in mammals) and ran across a quotation from Perry Metzger to the effect that so long as the laws of physics apply, there will always be evolution, hence nature red in tooth and claw will continue into the future—to him, the obvious analogy for the advent of AI was "nature red in tooth and claw," and people who see things this way tend to want to

> cling to that analogy even if you delve into some basic evolutionary biology with math to show how much it *isn't* like intelligent design. For Robin Hanson, the one true analogy is to the industrial . . . and farming revolutions, meaning that there will be lots of AIs in a highly competitive economic situation with standards of living tending toward the bare minimum, and this is so absolutely inevitable and consonant with The Way Things Should Be as to not be worth fighting at all. That's his one true analogy and I've never been able to persuade him otherwise. For Kurzweil, the fact that many different things proceed at a Moore's Law rate to the benefit of humanity means that all these things are destined to continue and converge into the future, also to the benefit of humanity. For him, "things that go by Moore's Law" is his favorite reference class.
>
> I can have a back-and-forth conversation with Nick Bostrom, who looks much more favorably on Oracle AI in general than I do, because we're *not* playing reference class tennis with "But surely that will be just like all the previous X-in-my-favorite-reference-class," nor saying, "But surely this is the inevitable trend of technology"; instead we lay out particular, "Suppose we do this?" and try to discuss how it will work, *not* with any added language about how surely anyone will do it that way, or how it's got to be like Z because all previous Y were like Z, *et cetera*.[1]

When we shared this blog, Eliezer and I had a long debate here on his "AI foom" claims. Later, we debated in person once. (See also slides 34–35 of this three-year-old talk.[2]) I don't accept the above as characterizing my position well. I've written up summaries before, but let me try again, this time trying to more directly address the above critique.

Eliezer basically claims that the ability of an AI to change its own mental architecture is such a potent advantage as to make it likely that a cheap, unnoticed, and initially low-ability AI (a mere "small project machine in a basement") could without warning, over a short time (e.g., a weekend) become so powerful as to be able to take over the world.

As this would be a sudden big sustainable increase in the overall growth rate in the broad capacity of the world economy, I do find it useful to compare this hypothesized future event to the other past events that produced similar outcomes, namely a big sudden sustainable global broad capacity-rate increase. The last three were the transitions to humans, farming, and industry.

I don't claim there is some hidden natural law requiring such events to have the same causal factors or structure, or to appear at particular times. But I do think these events suggest a useful, if weak, data-driven prior on the kinds of factors likely to induce such events, on the rate at which they occur, and on their accompanying inequality in gains. In particular, they tell us that such events are very rare, that over the last three events gains have been spread increasingly equally, and that these three events seem mainly due to better ways to share innovations.

Eliezer sees the essence of his scenario as being a change in the "basic" architecture of the world's best optimization process, and he sees the main prior examples of this as the origin of natural selection and the arrival of humans. He also sees his scenario as differing enough from the other studied growth scenarios as to make analogies to them of little use.

However, since most global bio or econ growth processes can be thought of as optimization processes, this comes down to his judgment on what counts as a "basic" structure change, and on how different such scenarios are from other scenarios. And in my judgment the right place to get and hone our intuitions about such things is our academic literature on global growth processes.

Economists have a big literature on processes by which large economies grow, increasing our overall capacities to achieve all the things we value. There are of course many other growth literatures, and some of these deal in growths of capacities, but these usually deal with far more limited systems. Of these many growth literatures, it is the economic growth literature that is closest to dealing with the broad capability growth posited in a fast-growing-AI scenario.

It is this rich literature that seems to me the right place to find and hone our categories for thinking about growing broadly capable systems. One should review many formal theoretical models, and many less formal applications of such models to particular empirical contexts, collecting data points of what is thought to increase or decrease growth of what in which contexts, and collecting useful categories for organizing such data points.

With such useful categories in hand, one can then go into a new scenario such as AI foom and have a reasonable basis for saying how similar that new scenario seems to old scenarios, which old scenarios it seems most like (if any), and which parts of that new scenario are central vs. peripheral. Yes, of course if this new area became mature it could also influence how we think about other scenarios.

But until we actually see substantial AI self-growth, most of the conceptual influence should go the other way. Relying instead pri-

marily on newly made-up categories and similarity maps between them, concepts and maps which have not been vetted or honed in dealing with real problems, seems to me a mistake. Yes, of course a new problem may require one to introduce some new concepts to describe it, but that is hardly the same as largely ignoring old concepts.

So I fully grant that the ability of AIs to intentionally change mind designs would be a new factor in the world, and it could make a difference for AI ability to self-improve. But while the history of growth over the last few million years has seen many dozens of factors come and go, or increase and decrease in importance, it has only seen three events in which overall growth rates greatly increased suddenly and sustainably. So the mere addition of one more factor seems unlikely to generate foom, unless our relevant categories for growth-causing factors suggest that this factor is unusually likely to have such an effect.

This is the sense in which I long ago warned against over-reliance on "unvetted" abstractions. I wasn't at all trying to claim there is one true analogy and all others are false. Instead, I argue for preferring to rely on abstractions, including categories and similarity maps, that have been found useful by a substantial intellectual community working on related problems. On the subject of an AI-growth foom, most of those abstractions should come from the field of economic growth.

* * *

See original post for all comments.

* * *

1. Eliezer Yudkowsky, "Reply to Holden on 'Tool AI,'" *Less Wrong* (blog), June 12, 2012, comment 71vj, http://lesswrong.com/lw/cze/reply_to_holden_on_tool_ai/.

2. Robin Hanson, "Economics of Nanotech and AI" (Paper presented at Foresight 2010: the Synergy of Molecular Manufacturing and AGI, January 16–17, 2010), slides 34–35 begin at 27m2s. Powerpoint file at http://hanson.gmu.edu/ppt/Econ of AI n Nanotech.ppt, http://vimeo.com/9508131.

61

AI-Foom Debate Summary

Kaj Sotala

28 January 2013

Editor's Note: This chapter contains many direct quotes from the preceding chapters, not all of which are marked as such. All ideas should be attributed to their original authors.

1. Introduction

An "intelligence explosion" is a hypothetical event in which a machine intelligence becomes better than humans at designing new machine intelligences,[1] potentially leading to a sequence of ever-more-intelligent machine intelligences that would leave humanity far behind. It has been proposed that humanity might become extinct as the result of such an event,[2] and that we should attempt to carefully design artificial intelligences in such a way that their values correspond to our own.[3]

In 2008, Robin Hanson and Eliezer Yudkowsky debated the possibility and consequences of an intelligence explosion on their blog, *Overcoming Bias*. They later held a ninety-minute debate on the issue in 2011. Eliezer Yudkowsky has been one of the main proponents of the need to develop safe, or "Friendly," artificial intelligences.[4] He founded and works at the Machine Intelligence Research Institute, which is dedicated to this goal. Robin Hanson is an economist at George Mason University and has published a number of papers on the societal and economic impacts of machine intelligence.[5] He expects a more decentralized and less threatening intelligence explosion, even though it could still be pretty fast compared to the economy's current growth rate. Hanson thinks that it is most likely to be caused by the capability to digitally emulate human brains, rather than by entirely new kinds of hand-coded artificial intelligence.

Hanson and Yudkowsky represent important positions on the intelligence explosion, and their conversations cover many arguments which have not yet been analyzed in the academic literature. Here we provide a summary of their debate.

2. Overview

In "Setting the Stage," Hanson establishes that both he and Yudkowsky agree upon the following points:

1. Machine intelligence would be a development of almost unprecedented impact and risk, well worth considering now.

2. Feasible approaches include direct hand-coding, based on a few big and lots of little insights, and on emulations of real human brains.

3. Machine intelligence will, more likely than not, appear within a century, even if the progress rate to date does not strongly suggest the next few decades.

4. Math and deep insights (especially probability) can be powerful relative to trend fitting and crude analogies.

5. Long-term historical trends are suggestive of future events, but not strongly so.

6. Some should be thinking about how to create "friendly" machine intelligences.

Hanson notes that the two disagree modestly on the chances of the emulation and direct-coding approaches, with Hanson considering the former, and Yudkowsky the latter, more likely to succeed first. However, the major disagreement is on "the chances that a single hand-coded [AI] will suddenly and without warning change from nearly powerless to overwhelmingly powerful." Hanson estimates the probability of this happening as less than 1%, while Yudkowsky puts the probability at more than 10%.

Yudkowsky's reasoning is based on the concept of *optimization power*, the general ability of a process to create specific situations that would have been very unlikely to emerge by random chance. Yudkowsky points out that the history of life on Earth so far has shown a trend toward processes with increasing optimization power. He presents theoretical arguments for why an artificial intelligence could be expected to rapidly obtain an enormous degree of optimization power relative to that of humanity.

Hanson is skeptical about the usefulness of the optimization power concept. He points out that academic studies on innovation and economic growth have produced models that have been tested in a variety of situations and over a long period of time. Hanson notes that, if we wish to make claims about a situation that has never happened before, we should use abstractions such as these, which some community has previously applied and found useful in understanding existing situations. In contrast, Yudkowsky's concept is based only on a handful of events, most of them so far away in time that it is hard to obtain much reliable information about them. While Hanson acknowledges that his models may be wrong, he considers them a much more robust tool for prediction than Yudkowsky's, and he does not expect any single player to achieve a position where they could quickly dominate all the others.

Hanson and Yudkowsky also disagree on the extent to which an AI's resources might be local as opposed to global, the extent to which knowledge is likely to be shared between various AIs, and whether an intelligence explosion should be framed as a "winner-take-all" scenario.

3. The Optimization Power Argument

3.1. Conceptual Background

In computer science, there is the notion of a "search space" or "solution space"—a conceptual space containing all the possible solution candidates for a problem. Different solutions can be said to be closer or further apart from each other. For example, if one is exploring car designs, then the design for a fifty-ton truck is closer to the design

of a forty-nine-ton truck than either is to the design of a sports car. Likewise, if one is trying to solve a problem such as which kind of car would be the fastest, the sports car is probably closer to the best solution than either of the trucks is. Depending on how the problem has been formalized, this distance can be measured in an objective manner.

Different problems may vary in the size of the search space, and in how easy it is to find a solution that actually solves the problem. For example, the problem "specify a molecule that is partially made up of carbon atoms" is much easier to solve than the problem "specify a configuration of atoms that's equivalent to a living cat."

We can say that the solutions to the "carbon atoms" problem make up a much larger fraction of the search space than the solutions to the "living cat" problem, as the relative number of goal states compared to the number of all possible states is larger. More explicitly, the fraction (all possible molecules with carbon)/(all possible molecules) is much larger than the fraction (all configurations of atoms which make up a living cat)/(all configurations of atoms).

If the region of solutions is large enough relative to the size of the search space, one may eventually find it with just a *blind search*. In our example, this would correspond to just picking various atoms at random, trying to fit them together, and then testing whether the produced molecule happens to fit our criteria—blindly jumping around the search space hoping to hit a solution by accident. If one is looking to come up with a molecule that has carbon atoms in it, one is likely to pick some carbon atoms and combine them in a valid way before too long. But if one wants to produce a living cat this way, the whole lifetime of the universe probably isn't enough.

If somebody has a complicated problem to solve, they need a more guided way of searching the space. For example, they might constrain themselves to a specific region—not just picking any atom at random, but always picking a carbon atom at first. Or they might come up with some measure of distance to their target, trying to always move in the direction that reduces the distance. Such an approach would be far more likely to find the right answer quickly than a mere blind search would be.

3.2. The Argument: Yudkowsky

Yudkowsky defines an *optimization process* as a process that hits very small targets in a very large search space. This can be either the space of possible futures, in which case we may talk about planning, or the space of possible designs, in which case we may talk about invention. Human intelligence is one example of an optimization process: human engineers reliably design artifacts such as cars that one would never find with a blind search. Even a very basic task like walking requires finding a narrow region in the space of all possible muscle movements: one wouldn't get anywhere by just randomly spasming their legs. Evolution is another example of an optimization process: it has created very unlikely creatures such as cats and humans. As the example of humans shows, some of evolution's creations are optimization processes themselves. If an optimization process is capable of hitting very improbable targets (relative to random selection) in a search space, it is said to have a lot of *optimization power*.

There's a straightforward analogy between optimization power and intelligence (as defined by Legg and Hutter[6]). Using their framework, take an agent that is deciding its actions at random. If the en-

vironment is complex and only very specific patterns of actions lead to high rewards, then that agent may have a very small probability of getting a high reward. In contrast, an intelligent agent has a much better chance of hitting the—*a priori* improbable—sequence of actions that produces a high reward. Furthermore, an intelligent agent may succeed in this in a great variety of different environments.

The analogy of evolution as an optimization process is somewhat imperfect, for a search in the computer science sense of the term implies an explicit goal, while evolution is just a process that happens, with no overarching goals. Nonetheless, evolution qualifies as an optimization process because it implements a *cumulative search* in a way that other physical processes, like star formation, do not. If one star burns brighter or longer, that does not affect the nature of the next star to form. There is only a blind search, with each star being picked more or less at random from the space of possible stars. The probability of seeing a star at any given point of space is given by the probability that a star will form multiplied by the average duration of a star.

Analysis. It feels like this should be made more rigorous, or otherwise explained better. One could argue that star formation *is* a cumulative search in the sense that the current state of the universe affects future states: most stars do not simply pop out of pure vacuum, Boltzmann-brain-like, but are instead formed out of existing matter by a gradual process. It would also have been very unlikely for our current galaxy to simply materialize into existence right off the bat. Instead it came to be by a process of galaxy formation that searched the space of possible galaxies and eventually hit this point.

Optimization processes were introduced to Earth with the first replicator. Perhaps the probability that a single replicator would form was 10^{-30}, and perhaps it made 10,000,000,000 copies of itself. If you were observing things at random, not just on Earth but on all the planets with tidal pools, this would increase your probability of encountering a replicator by a factor of 10^{10}, with the total probability going up to 10^{-20}.

More importantly, the copying process was not perfect, so some of the copies were different from the original. Some of those changes helped the replicators survive or to replicate themselves better, and such replicators increased their numbers. This was an optimization process in the sense that the first replicator explored the neighboring regions of the search space—some of which contained replicators better capable of surviving and copying themselves. After such better replicators had been created, they explored *their* neighborhoods, again eventually leading to the creation of yet better replicators. The probability of seeing such better replicators, if looking randomly at all the planets in the universe, began to increase. Eventually, life took over the whole planet.

In studying optimization processes, Yudkowsky wishes to separate the meta level from the object level. In other words, to separate the structure of the optimization being performed from that which is being optimized. In evolution, the meta level consists of things such as sexual recombination and natural selection on asexual populations. The object level consists of things such as trees, butterflies, and humans. The object level is far more complicated than the meta level. This is because the meta level is something that accidentally be-

gan to happen one day, while the object level is the end result of a long process of optimization.

At different times, a tiny number of seemingly trivial innovations, like bundling different genes together, separating information storage from moving machinery, and randomly recombining groups of genes, fed back from the replicators to the meta level. These meta-level changes increased evolution's optimization power enough that biologists consider them to structure the evolutionary epochs of life on Earth. However, the core process of evolution still remains very simple, even though it has been capable of producing immensely complex object-level outcomes.

Evolution does feed on itself in the sense that each new adaptation opens up new avenues of further adaptation, but this happens almost entirely on the object level: the development of the first light-sensitive cells made possible the later development of eyes. The meta level mostly operates under the same rules as it always has.

The first animal brains had some optimization power—they could (literally) search their environment. But for the most part, animal brains were things that evolution optimized, not things that would have exerted considerable optimization power on their own. A cat's brain obtains knowledge over a lifetime, but eventually the cat dies and the knowledge is lost instead of accumulating. Compared to evolution, animal brains lacked *cumulative optimization power*, as their products did not accumulate complexity over time. They also lacked *generality of optimization power*, as they could not produce the vast range of artifacts produced by evolution.

Humans, on the other hand, exert quite a lot of optimization power. While natural selection takes hundreds of generations to do

anything and millions of years to create new complex designs, human programmers can design a complex machine with a hundred interdependent elements in a single afternoon. Natural selection is an accidental optimization process, while humans are *optimized* optimizers.

A human engineer—drawing on the accumulated knowledge and skill of other humans—can in a short time come up with designs that the whole of evolution could *never* have developed. This is despite the fact that humanity's biomass is a miniscule proportion of all the biomass on Earth. The amount of resources that can be put into searching the space matters much less than the *efficiency* of the search: humanity, despite having far less resources, is far more efficient in using them.

Thus we can infer at least two components of the *optimization velocity* of a process:

- The *optimization resources*, like the amount of computing power available to a fixed program, or the number of individuals in a population pool.
- The *optimization efficiency*, the relation between resources invested and search power generated, which is presumably a function of the optimizer's structure at that point in time.

Also sometimes we are closer or farther away from the solution, or a solution may be harder to reach. This gives us the third component:

- The searchability of the neighborhood of the current location, and the availability of good/better alternatives in that rough region. Call this the *optimization slope*. Are the fruit low-hanging or high-hanging, and how large are they?

Distance isn't just a degree of similarity. In biology, different mutations have different probabilities of appearing in future generations, depending on the fitness benefit (or penalty) that they confer on an organism. Suppose that there are two different chains of mutations: chain A, which is three mutations long, and chain B, which is six mutations long. Now, although the outcome of the first chain of mutations can be said to be *closer* in the search space, it might be that each mutation in the second chain confers a much greater fitness advantage, thus having a higher chance of spreading in the population once they come into existence. Thus the *optimization slope* is more slanted toward the solution of the second chain.

So far, most of the optimizing has been done by natural selection: a process of beings imperfectly replicating themselves and some of them surviving better than others. This process has been exerting a relatively constant optimization pressure: its optimization resources have grown, but for the most part, its optimization efficiency has not. There have been some exceptions, such as the emergence of cells and DNA. These have increased evolution's optimization efficiency to such an extent that they're considered major evolutionary milestones.

4. Recursive Self-Improvement

Yudkowsky discusses the concepts of cascades, cycles, insight, and recursion:

Cascades are when one development leads to another. It's hard to know what happened to separate us from chimps, but regardless,

the difference between humans and chimps isn't just *one* change, but rather a cascade of them that never got started in our closest relatives.

Cycles are when optimization A benefits optimization B, which then benefits A again. They can be thought of as repeatable cascades that happen with a high regularity. The development of writing increased the speed by which humanity accumulated discoveries, but improvements to writing itself were relatively rare—once writing had been discovered, that discovery could not simply be repeated over and over to gain a boost on each time. As an example of a cycle, Yudkowsky uses the example of a self-sustaining nuclear reaction in physics. The key number for a pile of uranium is k, the effective neutron multiplication factor—the average number of neutrons from a fission reaction that go on to cause another fission reaction. At $k < 1$, the pile is subcritical. At $k \geq 1$, the pile will sustain a critical reaction, each fission creating, on average, at least one more fission. Another important cycle is compound interest on investment, where the interest that has been added to the initial investment earns additional interest.

Insight is when some piece of knowledge vastly increases one's optimization efficiency by making it easier to search the space. An insight is a chunk of knowledge which, if one possesses it, decreases the cost of solving a whole range of governed problems. Calculus and algebra, for example, make many kinds of math problems drastically easier to solve. It is the difference between evolution "nibbling bits off the immediate search neighborhood" and the human ability to jump straight to the right answer. An insight consists of understanding what's "good" about an idea in a way that divorces it from any single point in the search space. Some examples are the insight of calcu-

lus apart from gravity, the insight of mathematical physics apart from calculus, and the insight of math apart from mathematical physics.

Recursion is when an optimization process can improve itself *directly* and these improvements make it more efficient to create further changes. Evolution has so far only been very weakly recursive: it has come up with discoveries that made it faster, but there has been a long delay between these changes, and they haven't affected the *core* process—of organisms being selected on the basis of their differential ability to replicate and survive.

Natural selection seems to have produced a pretty smooth trajectory of more sophisticated brains over the course of hundreds of millions of years. Thus:

- Natural selection on sexual multicellular eukaryotic life can be treated, to a first-order approximation, as an optimizer of *roughly constant efficiency and constant resources*.
- Natural selection does not have anything akin to insights. It does sometimes stumble over adaptations that prove to be surprisingly reusable outside the context for which they were adapted, but it doesn't fly through the search space like a human. Natural selection is just *searching the immediate neighborhood of its present point in the solution space, over and over and over*.
- Natural selection *does* have cascades: adaptations open up the way for further adaptations.

Yudkowsky admits that there is debate over whether or not the evolution of biological brains has accelerated, but argues that the speed

of evolution does not seem to be logarithmic or decelerating. With constant optimization pressure from natural selection, and no intelligent insight, there were no diminishing returns to a search for better brain designs up to at least the human level, and there were probably accelerating returns.

For example, it did *not* take ten times as long to go from *H. erectus* to *H. sapiens* as from *H. habilis* to *H. erectus*. Hominid evolution did *not* take eight hundred million years of additional time to produce humans, after evolution immediately produced *Australopithecus*-level brains in just a few million years after the invention of neurons themselves. Human intelligence does *not* require a hundred times as much computing power as chimpanzee intelligence. Human brains are merely three times too large, and our prefrontal cortices six times too large, for a primate with our body size. It does not seem to require a thousand times as many genes to build a human brain as to build a chimpanzee brain, even though human brains can build toys that are a thousand times as neat.

Yudkowsky suggests the following hierarchy of causality for an intelligent mind:

- **The metacognitive level** is the original optimization process that builds the mind. In the case of a human, this refers to natural selection. In the case of an AI, this either refers to human programmers, or, after some point, to the AI itself.

- **The cognitive level** is built by the metacognitive level. In humans, this refers to the labor performed by one's neural circuitry, algorithms that consume large amounts of computing

power but are mostly opaque to a person. You know what you're seeing, but you don't know how the visual cortex works.

- **The metaknowledge level** consists of discoveries about how to discover. "Science" is an archetypal example. This can be thought of as reflective cognitive content (knowledge about how to think). Metaknowledge can be conveyed and accumulated across generations; centuries later, we still remember how to do science.
- **The knowledge level** consists of knowledge about various things in the world—for example, knowing how gravity works.
- **The object level** involves specific actual problems, like building a bridge.

An AI programmer, asked to write a program that plays chess, will tackle the task using their existing knowledge and insight in the domain of chess and search trees; they will apply any metaknowledge they have about how to solve programming problems or AI problems; they will process this knowledge using the deep algorithms of their neural circuitry; and this neutral circuitry will have been designed (or rather its wiring algorithm designed) by natural selection.

An AI, asked to write a program that plays chess, might do the same thing. It would use its knowledge, metaknowledge, and existing cognitive algorithms. The difference is that the AI's metacognitive level is not natural selection, but the object level of the programmer who wrote the AI, using their knowledge and so on.

An AI might also be asked to write a better algorithm than X for storing, associating to, and retrieving memories. In one sense, this

is just another object-level problem. But if the AI itself uses algorithm X to store associative memories, then if the AI can improve on this algorithm, it can rewrite its code to use the new algorithm X+1. This means that the AI's metacognitive level—the optimization process responsible for structuring the AI's cognitive algorithms in the first place—has now *collapsed to identity* with the AI's object level.

This is different from the ordinary kind of improvement process that humanity is undergoing. While it has long been possible for humans to experiment with various ways of improving themselves, they have never had the ability to *directly* see and modify their neural circuitry. The fact that humans do not yet understand their neural circuitry is the reason why they have not yet created an AI.

Evolution is not *recursive* in the sense of evolution's discoveries being used to make the process of evolution itself faster or more effective. While sometimes evolution stumbles upon improvements that accelerate it, this is not a systematic trend or an explicit goal. There's no strong link between evolution's object-level discoveries and the mechanism by which evolution operates. Despite this, it has been able to produce better brains at an accelerating, or at least linear, rate. A strongly recursive AI, with its object level being directly linked to its metacognitive level, could plausibly make far faster progress.

So far, the metacognitive level (natural selection) has been exerting a roughly constant pressure to improve the cognitive level (human intellect), which has over the narrower domain of recorded history been exerting a roughly constant pressure to improve the metaknowledge level (professional specialization, science, etc.), which has been exerting an increasing pressure to improve the knowledge level (all our accumulated knowledge), which has been exerting an increasing

pressure to improve the object level. With self-improving AI, the end result of all the optimization pressure on the object level feeds back into the metacognitive level, which has never happened before.

As a rough general analogy, the impact of recursion could be described as replacing the equation $y = f(t)$ with $\frac{dy}{dt} = f(y)$. For example, if somebody had bought a bond and they spent the earned money every year (instead of reinvesting it), their total interest over time would be a linear $y = m \cdot t$. If they instead reinvested it, the return would become $\frac{dy}{dt} = m \cdot y$, with the solution $y = e^{(m \cdot t)}$. While Yudkowsky does not believe that one could solve similar equations to get a description of the growth rate of a self-improving AI, he does think that it's a reason why the future isn't well described by past trends—because it contains a feedback loop that the past doesn't.

Now, it's not a given that this would lead to very fast progress—it might also lead to zero progress.

Optimizing compilers are programs designed to make computer programs faster by introducing improvements to the way the code is written and by eliminating unnecessary processing steps. An optimizing compiler set to improve itself will produce a single series of improvements, making itself slightly faster. After that, the compiler has already performed all the improvements that it can—it cannot further improve itself to make itself even faster.

The self-improving EURISKO AI system employed heuristics in order to solve problems in a variety of domains. It also had heuristics for suggesting new heuristics, and metaheuristics could apply to any heuristic, including metaheuristics. For example, EURISKO started with the heuristic "investigate extreme cases" but moved on to "investigate cases close to extremes." It could even modify the metaheuris-

tics that modified heuristics. Yet, after a while, it could no longer find useful modifications. Its self-improvements did not spark a sufficient number of new self-improvements. Eurisko did not start out with human-level intelligence plus the ability to modify itself—its self-modifications were produced by the simple procedural rules of some heuristic or other.

Yudkowsky claims that a self-improving AI should "either flatline or blow up." There exists a great range of potential self-improvement speeds, of which only a very narrow part would look like gradual improvement to humans. It would take exactly the right law of diminishing returns to hit the range where humans could see the AI making progress, but not so fast that humans couldn't keep up.

5. *Hard Takeoff*

According to Yudkowsky, an AI engaging in recursive self-improvement might undergo "hard takeoff," an event where it rapidly gains enough power and intelligence to become the dominant force on Earth. But even without presuming explosive recursive self-improvement, there may very well be a hard takeoff. The advent of human intelligence was a discontinuity even without recursive self-improvement.

The differences between humans and chimps are relatively minor—both species have similar brain architectures divided into frontal cortex, cerebellum, etc.—suggesting that only a small amount of improvement sufficed to create human-level intelligence from chimp intelligence. While Yudkowsky admits this is only suggestive evidence, it lightly suggests and provides a hypothetical illustration

of a discontinuous leap upward in capability that results from a relatively small amount of improvement. There may equivalently be similar points for AIs, allowing considerably better solutions than before as a result of a few final tweaks to the mind design.

Another way of undergoing a hard takeoff is simply acquiring more computational resources. An AI might be improving itself, but doing it at a very slow rate. If it is upgraded to a much more powerful system, this could speed up its research.

With a sufficiently stupid algorithm, a few orders of magnitude more computing power would only mean a linear increase in performance. On the other hand, smarter algorithms might benefit more. Humans have a brain three times as large, and a prefrontal cortex six times as large, as that of a standard primate our size, suggesting that an exponential improvement in resources isn't needed for a linear improvement. Yudkowsky admits that this analogy may not be correct, in that humans might not have much more horsepower than chimps, but merely take better advantage of it. But evolution does suggest that minds do not run into sharply diminishing returns on processing power in the course of reaching human intelligence, even when the processing power increase is strictly parallel rather than serial.

If the AI obtains (for instance) a ten-thousand-fold increase in its computing resources, all future improvements will now have ten thousand times as much computing power available. A single improvement to code now has more impact than before, and is liable to produce more improvements. Recalling the uranium pile analogy, the pile is always running the same "algorithm" with respect to neutrons causing fissions that produce further neutrons. Yet piling on more uranium can cause it to go from subcritical to supercritical, as

any given neutron has more uranium to travel through and a higher chance of causing future fissions.

One way of acquiring more resources is to simply wait and allow better hardware to be developed. Another would be the discovery of a way to take over all the poorly defended computers on the Internet. Yudkowsky suggests that this may not require what humans would regard as genius, just the ability to examine lots of machine code and do relatively low-grade reasoning on millions of bytes of it.

Another kind of resource hardware boost would be represented by modern CPUs having a 2 GHz *serial* speed, in contrast to neurons that spike a hundred times per second. The "hundred-step rule" in computational neuroscience is a rule of thumb that any postulated neural algorithm which runs in real time has to perform its job in less than a hundred serial steps one after the other. Much of the brain's parallelism could consist of cache lookups to make up for the brain's serial slowness. A correctly designed midsize computer cluster might be able to get high-grade thinking done at a serial speed much faster than human, even if the total parallel computing power was less.

The development of an AI should also be expected to hit a discontinuity at the point where the AI obtains insight into its own workings: the point where it could not only generate its own source code, but also write rewrite a major AI textbook on its own. At this point, the AI will become capable of contributing to its development, and AI research will likely accelerate quickly.

Yudkowsky says that his analysis permits at least three possible AI trajectories:

1. An AI is created by researchers who are good at finding tricks that work, but who have at most a partial insight to the way a mind works. The AI is less intelligent than the researchers, but performs lower-quality operations much faster. This mind finds a set of mutually supporting self-improvements, cascades up to the level of a very smart human, achieves insight into intelligence, and rapidly improves itself to superintelligence.

2. Researchers with partial insight create a mind that performs a number of tasks very well, but can't handle self-modification let alone AI theory. A mind like this might progress with something like smoothness, pushed along by the researchers rather than itself, even all the way up to average-human capability, not having the insight into its own workings to push itself any further. We also suppose that the mind either is already using huge amounts of available hardware or scales *very* poorly, so it cannot undergo hard takeoff by simply adding hardware. Yudkowsky thinks this scenario is less likely, but that it is not *ruled out* by any effect he can see.

3. Researchers with strong insight into intelligence create a mind capable of modifying itself with deterministic precision—provably correct or provably noncatastrophic self-modifications. Yudkowsky considers this the only plausible path to Friendly AI.

Yudkowsky's analysis does not permit a scenario where an AI undergoes a cycle of self-improvement, starting from stupidity, that carries it up to the level of a very smart human and then stops, unable

to progress any further. Neither does it seem to permit a scenario where an AI is pushed by its programmers from a roughly human level to the level of a very smart human to a mildly superhuman level, but the mind still does not achieve insight into its own workings and still does not undergo an intelligence explosion—just continues to increase smoothly in intelligence from there.

6. Questioning Optimization Power

6.1. The Issue of Abstractions

An *abstraction* is a model that neglects some details to emphasize others; the right choice of an abstraction depends on what one wants to do. Yudkowsky's optimization power concept is one kind of abstraction. However, Hanson, whose background is in economics, prefers the abstractions developed in the academic studies of innovation and economic growth, finding them more relevant and better tested in a wide variety of situations. Applying these abstractions, the most relevant major transitions have been developments like farming and industry.

Hanson's models do not predict a rapid takeover by a single entity. Rather, they predict that development will be interdependent and gradual, with most innovations becoming broadly dispersed between many different actors.

Hanson is skeptical about the recursive self-improvement and hard takeoff scenarios, saying that, while it's easy to think of ways by which AI development could be considered "recursive," standard growth theory already has many examples like it. For example, a rise in population provides more people to develop innovations of

all sorts; lower transportation costs allow more scale economies over larger integrated regions for many industries; tougher equipment allows more areas to be farmed, mined, and colonized; and lower information storage costs allow more kinds of business processes to be studied, tracked, and rewarded. None of this has historically led to a single entity taking over the world.

Hanson argues that if you wish to use some sort of abstraction, you should try to test it in as many situations as possible. He writes: "If you came up with an account of the cognitive processes that allowed Newton or Einstein to make their great leaps of insight, you would want to look for where that, or related accounts, applied to more common insight situations. An account that only applied to a few extreme "geniuses" would be much harder to explore, since we know so little about those few extreme cases. . . . 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."

Yudkowsky replies that the economical research that Hanson relies on does not model cognitive phenomena. All of this research has been documenting humans with human brains, and all of these models and the experiments made to test them have assumed human minds. When the assumptions made in the economical growth literature fail to apply, the models break down. While economics does have papers about cognitive phenomena, they're dealt with on a very superficial level. For example, a seminal paper in the endoge-

nous growth literature, which tries to study the generation of ideas, talks about ideas being generated by combining other ideas, so that if you've got N ideas already and you're combining them three at a time, that's a potential $N!/((3!)(N-3!))$ new ideas to explore, a claim with little empirical backing and which seems too specific for the model. It talks about ideas in the economy, not about an economy of ideas.

Yudkowsky thinks that the standard economic models incorrectly assume that scientific research and economic growth will continue to be carried out by essentially unmodified human minds, with the same cognitive capabilities as today's humans. He writes: "Would the history of the world *really* be just the same, proceeding on *just exactly* the same timeline as the planets move in their orbits, if, for these last fifty years, the researchers themselves had been running on the latest generation of computer chip at any given point? That sounds to me even sillier than having a financial model in which there's no way to ask what happens if real estate prices go down."

Hanson points out that all models have some unrealistic aspects. We can't conclude from the fact that a seminal model has some unrealistic aspects that it is useless or that an almost *entirely* unvetted concept (such as Yudkowsky's optimization power concept), which is also likely to contain some unrealistic aspects, would do better. As for the claim that economics assumes human minds, the standard model mind used in economics is an expected utility maximizer.

Yudkowsky comments that simply saying, "Your abstractions are not vetted," makes it hard for him to reply properly. While he admits that Hanson's point against unvetted abstractions is a strong one, it nonetheless seems wrong to prefer a model that treats human brains as black boxes which are never opened and improved upon. In the

standard model, the brain is never made bigger or faster or has its software redesigned. While the lack of vetted abstractions makes the problem harder to analyze, the fact that so many normal assumptions break down is why one should regardless *try* to analyze it. Yudkowsky's core argument is about what happens when one does pry apart the black boxes—if one rejects all such speculation as "unvetted abstraction," it doesn't leave much to talk about.

Hanson replies that he's not saying no one should analyze the assumptions of changing brains, he's saying that we should prefer to do such analysis with vetted abstractions.

Hanson references his earlier paper,[7] an economic growth model that deals with machine intelligences that can be copied or sped up. In economics, the simplest standard model of endogenous growth is "learning by doing," where productivity increases with practice. Hanson used this approach to model Moore's Law and faster ems (whole-brain emulations) in his paper. He also notes that "while economists have many abstractions for modeling details of labor teams and labor markets, our standard is that the simplest versions should be of just a single aggregate quantity of labor. This one parameter of course implicitly combines the number of workers, the number of hours each works, how fast each thinks, how well trained they are, etc. If you instead have a one-parameter model that only considers how fast each worker thinks, you must be implicitly assuming all these other contributions stay constant. When you have only a single parameter for a sector in a model, it is best if that single parameter is an aggregate intended to describe that entire sector, rather than a parameter of one aspect of that sector."

Yudkowsky: "If one woman can have a baby in nine months, nine women can have a baby in one month? Having a hundred times as many people does not seem to scale even close to the same way as the effect of working for a hundred times as many years. This is a thoroughly vetted truth in the field of software management." Yudkowsky does not consider Hanson's model well vetted either, and is skeptical about what makes Hanson's extensions of economic theory vetted while his concepts aren't.

6.2. The Historical Record

Hanson would like to see the optimization power model tested better. He notes that, on a rough level, Yudkowsky seems to be essentially positing a three-level hierarchy:

1. The dominant optimization process: natural selection, brains with culture, or full AI
2. Improvements that aid that process, such as cells, sex, writing, or science
3. Key "object-level" innovations that open the path for other such innovations

Hanson describes the major developments in the traditional fossil record as "Any Cells, Filamentous Prokaryotes, Unicellular Eukaryotes, Sexual(?) Eukaryotes, and Metazoans," and notes that perhaps two of these five events are at Yudkowsky's level two, and none at level one. Relative to these events, the first introduction of human culture isn't remotely as noticeable. While the poor fossil record means we shouldn't expect a strong correspondence between the biggest inno-

vations and dramatic fossil events, we can at least say this data doesn't strongly support Yudkowsky's ranking.

Our more recent data is better, allowing clearer tests. The last three strong transitions were humans, farming, and industry, and in terms of growth rate changes these seem to be of similar magnitude. Yudkowsky seems to predict we will discover the first of these was much stronger than the other two. And while the key causes of these transitions have long been hotly disputed, with many theories in play, Yudkowsky seems to pick specific winners for these disputes: intergenerational culture, writing, and scientific thinking. This seems wrong. While the introduction of writing did roughly correspond in time with farming it just doesn't seem plausible that writing caused farming, rather than vice versa. Few could write and what they wrote didn't help farming much. Farming seems more plausibly to have resulted from a scale effect in the accumulation of innovations in abilities to manage plants and animals—we finally knew enough to be able to live off the plants near one place, instead of having to constantly wander to new places.

For industry, the key innovation does not seem to have been a scientific way of thinking—that popped up periodically in many times and places, and by itself wasn't particularly useful. Hanson's guess is that the key was the formation of networks of science-like specialists, which wasn't possible until the previous economy had reached a critical scale and density.

Yudkowsky's response is that it may not be easy to discover the speed of development from the historical record. He is trying to measure the optimization velocity of information, not production or growth rates. Although this will translate into power eventually, mea-

suring things like the amount of biomass in the world may not reveal much about the optimization pressure.

For example, if there are fixed resources available then any evolutionary "progress" that we would recognize as producing a better-designed organism may just result in the displacement of the old allele by the new allele—*not* any increase in the population as a whole. It's quite possible to have a new wolf that expends 10% more energy per day to be 20% better at hunting, in which case the sustainable wolf population will decrease as new wolves replace the old." We shouldn't surprised if we have difficulty actually *observing* evolution speeding up with the advent of, e.g., sex, though it still seems to have happened.

Hanson notes that if Yudkowsky can't connect his theories to the historical record, there's little reason to believe in them. Yudkowsky answers that the amount of evidence he needs for his theory will depend on the strength of the predictions he wants to make with it. He would need far more evidence if he were to predict the specific speed at which an AI might obtain optimization power, but that is the reason why he is sticking to rough, qualitative predictions. His main prediction is that an AI's development trajectory will not look like a smooth, gradual development to us.

Yudkowsky also suggests that there are three valid ways of making predictions:

- Some problem domains are sufficiently well-understood to be precisely predictable. In such a domain, human knowledge can be used to exactly predict even kinds of outcomes that have never been seen before. For example, using the known laws of physics, one could plot the trajectory of the first moon rocket

before it was ever launched, or verify a computer chip before it is ever manufactured.

- Other problem domains are less well understood, hard to model exactly, and tend to run into unforeseen complications. In these cases, it is often best to take what is known as the "outside view" and predict that this event will happen roughly the same way as previous events of a similar kind.

- Some domains are even more novel, as they genuinely involve entirely new kinds of events that have never been seen before. In these cases, the outside view does not work, as there is no history of similar cases to compare with. In that case, the only thing that can be done is to apply a "weak inside view." This involves trying to model the causal process and producing "loose, qualitative conclusions" about only those issues where there seems to be lopsided support.

Yudkowsky considers the creation of AI to be a kind of event that has never been seen before, and where all attempts at offering precise quantitative predictions fail. Instead, he thinks that, looking at causal factors that have historically made various optimization processes powerful, one ought to make the "loose, qualitative" prediction that an AI is likely to become more powerful very quickly in human terms. Saying exactly *how* quickly isn't something that could be done, however.

Hanson is also skeptical about Yudkowsky's claim that natural selection has been exerting a relatively constant optimization pressure, or that its optimization efficiency has remained roughly stable. A

"smooth" trajectory could be caused by a constant as well as a non-constant efficiency, and the ways that genes get organized might enable evolution to search and reuse abstractions. The slow collection of a library of design parts may plausibly have been increasing evolution's optimization efficiency. And while new species do show up at a roughly constant rate, without some measure of how much better some species were than others, this doesn't imply a constant rate of improvement in something important.

Hanson thinks that we already understand the key difference between humans and chimps: an ability to save and accumulate knowledge that was previously lost with death. So the question is whether we can see a similar future gain: something that is now continually lost that would instead be allowed to accumulate.

6.3. The UberTool *Question*

Hanson introduces the thought experiment of *UberTool*, a company which "claimed that it had identified a set of mutually improving tools, sparking off a continuous sequence of self-improvement until the company could eventually come to dominate most industries in the world." He notes that such claims would not seem very plausible for most people.

Hanson finds a historical *UberTool* candidate in Douglas Engelbart, who in 1962 attempted to create a set of tools for improving the human intellect. While Engelbart's ideas had important legacies, he lost most of his funding in the early 1970s and his team dispersed. Even though Engelbart understood key elements of tools that today greatly improve team productivity, his team was not radically more productive, even at the task of improving their tools.

Hanson elaborates: "The point is that most tools require lots more than a few key insights to be effective—they also require thousands of small insights that usually accumulate from a large community of tool builders and users." Although there have been times when small teams have suddenly acquired disproportionate power, Hanson can't think of any time when such sudden small team power came from an *UberTool* scenario of rapidly mutually improving tools. He asks, why would one consider such an AI scenario plausible, if one doesn't consider such an *UberTool* scenario plausible? Why would a self-improving AI be so much more autonomous than a self-improving tool team?

Yudkowsky's response is that Engelbart was *insufficiently recursive*. Yudkowsky's concepts are about "strong recursion"—where the recursion feeds into whatever factor it is that determines most of the performance, improving it enough to make it possible to come up with further improvements. If A improves B by 50%, and B makes up 5% of A's performance, then A making this improvement to B improves A by 2.5%, which may not be enough to find further improvements and continue the self-improvement process. In contrast, if B makes up half of A's performance, then the improvement will be 25%, which has a much larger chance of yielding extra improvements.

Most of what the human brain does happens below the level of conscious notice, and although innovations like copying and pasting do reduce the amount of time needed to fight the typewriter, only a small part of the intellectual labor actually goes into fighting the typewriter. Engelbart could help one to copy and paste more easily, but he could not rewrite the hidden portions of the brain that labor to come up with good sentences and good arguments. The improve-

ment in efficiency could not be usefully reinvested to further improve efficiency—to do that properly would have required the ability to improve the brain itself.

It takes too much *human* labor to develop computer software and computer hardware, and this labor cannot be automated away as a one-time cost. If the world outside one's window has a thousand times as many brains, a 50% productivity boost that only cascades to a 10% and then a 1% additional productivity boost will not let one win against the world. If one's *UberTool* was itself a mind, if cascades of self-improvement could fully automate away more and more of the *intellectual* labor performed by the outside world—then it would be a different story. For as long as the development path requires thousands and millions of engineers and one can't divert that path through an internal computer, one is not likely to pull far ahead of the world. One can just choose between giving one's own people a 10% boost, or selling one's product on the market to give lots of people a 10% boost.

If one is getting most of one's technological progress *handed to one*—one's resources not being sufficient to do it in-house—then one won't be able to apply one's private productivity improvements to most of one's actual velocity, since most of one's actual velocity will come from outside. If one only creates 1% of the progress that one uses, then a 50% improvement becomes a 0.5% improvement. The domain of potential recursion and potential cascades is much smaller, diminishing k.

One might think that the development of computers is already recursive, since hardware engineers use better computers to develop yet better computers. But this recursion is weak compared to a scenario where researchers themselves run on computers. As a thought exper-

iment, giving researchers a computer twice as fast to analyze chips on would have less impact than a computer that made the researchers themselves run twice as fast.

Hanson thinks that a model which concentrates only on the speed at which researchers run is extremely stark, and leaves out various other variables that are usually taken into account in even the simplest standard growth models. The economy already has many loops of mutually reinforcing growth factors that do not result in accelerating growth.

7. Hanson's Intelligence Explosion Scenario

Hanson believes that whole-brain emulations are more likely to succeed in the near term than hand-crafted AIs. While the development of emulations would cause considerable economic change, it's not fundamentally different from previous economic breakthroughs, such as the Industrial Revolution. Hanson:

> Eventually, however, a project would succeed in making an emulation that is clearly sane and cooperative. . . . But enormous investment would be attracted to this race once news got out about even a very expensive successful emulation. As I can't imagine that many different emulation approaches, it is hard to see how the lead project could be much more than a year ahead. . . .
>
> Some project would start selling bots when their bot cost fell substantially below the (speedup-adjusted) wages of a profession with humans available to scan. Even if this risked more leaks, the vast revenue would likely be irresistible. This revenue might help this group pull ahead,

> but this product will not be accepted in the marketplace overnight. It may take months or years to gain regulatory approval, to see how to sell it right, and then for people to accept bots into their worlds, and to reorganize those worlds to accommodate bots. . . .
>
> In the absence of a strong world government or a powerful cartel, it is hard to see how the leader could be so far ahead of its nearest competitors as to "take over the world." Sure the leader might make many trillions more in profits, so enriching shareholders and local residents as to make Bill Gates look like a tribal chief proud of having more feathers in his cap. A leading nation might even go so far as to dominate the world as much as Britain, the origin of the Industrial Revolution, once did. But the rich and powerful would at least be discouraged from capricious devastation the same way they have always been, by self-interest.

Hanson argues that historical data says the inequality caused by major transitions is decreasing. The transition to multicellular organisms caused huge inequality, in that a probably tiny initial lineage soon came to dominate the energy usage, if not the mass, of life on Earth. The development of human brains likewise led to huge inequality as, whatever the size of the first species or lineage to embody the key brain innovation, later bottlenecks led to at most a few thousand individuals giving rise to all the individuals now. For the lineages that first mastered farming, the advantage was less overwhelming: In Europe,[8] Africa,[9] and Bali it seems post-transition population was about 20–50% from invading farmer groups, and the rest from the previous locals. Locals learned to adapt invader techniques fast enough to survive.

For the Industrial Revolution, the advantage seems even smaller. In 1500, Western Europe seems to have had about 18% of world population,[10] and today it has about 4%.[11] It seems unlikely that more than half of people today are descended from year-1500 Western Europeans. So they seem to have gained less than a relative factor of 2.5 in number of descendants by starting the Industrial Revolution. In GDP terms they have gained more of course.

Edinburgh gained some advantage by being the beginning of the Industrial Revolution, but it didn't take over the world. Northern Europe got closer to that goal, but still didn't take over the world. Various cities and countries needed each other and a large economy.

Hanson offers three reasons why the advantages accruing to early adopters are decreasing:

1. The number of generations per population doubling time has decreased,[12] leading to less inequality per doubling time. So if the "first mover's advantage" lasts some fixed number of doubling times before others find similar innovations, that advantage persists for fewer generations.

2. When lineages cannot share information, then the main way the future can reflect a new insight is via insight holders displacing others. As we get better at sharing info in other ways, the first insight holders displace others less.

3. Independent competitors can more easily displace each another than interdependent ones. For example, although it started the Industrial Revolution, Britain did not gain much relative to the rest of Western Europe; Western Europe as a

> whole gained much more relative to outsiders.[13] So as the world becomes interdependent on larger scales, smaller groups find it harder to displace others.[14]

Hanson points out that the first contribution is sensitive to changes in generation times, but the other two come from relatively robust trends. An outside view thus suggests only a moderate amount of inequality in the next major transition—nothing like a basement AI taking over the world.

Hanson also notes a number of factors that influence the variance in the outcome of an economic competition. The larger the variance, the better the best firm will do relative to the average, second best, or worst. His argument can be read to imply that an analysis based merely on "optimization power" ignores many of these factors, though Yudkowsky does not disagree with the list.

1. **Resource Variance:** The more competitors vary in resources, the more performance varies.

2. **Cumulative Advantage:** The more prior wins help one win again, the more resources vary.

3. **Grab It First:** If the cost to grab and defend a resource is much less than its value, the first to grab can gain a further advantage.

4. **Competitor Count:** With more competitors, the best exceeds the second best less, but exceeds the average more.

5. **Competitor Effort:** The longer competitors work before their performance is scored, or the more resources they spend, the more scores vary.

6. **Lumpy Design:** The more quality depends on a few crucial choices, relative to many small choices, the more quality varies.
7. **Interdependence:** When firms need inputs from each other, winner gains are also supplier gains, reducing variance.
8. **Info Leaks:** The more info competitors can gain about others' efforts, the more the best will be copied, reducing variance.
9. **Shared Standards:** Competitors sharing more standards and design features, in info, process, or product, can better understand and use info leaks.
10. **Legal Barriers:** May prevent competitors from sharing standards, info, inputs.
11. **Anti-Trust:** Social coordination may prevent too much winning by a few.
12. **Sharing Deals:** If firms own big shares in each other, or form a coop, or just share values, may mind less if others win. Lets tolerate more variance, but also share more info.
13. **Niche Density:** When each competitor can adapt to a different niche, they may all survive.
14. **Quality Sensitivity:** Demand/success may be very sensitive, or not very sensitive, to quality.
15. **Network Effects:** Users may prefer to use the same product regardless of its quality.

Hanson argues that if one worries about one competitor severely dominating all the others, one should attempt to promote factors that reduce success variance.

8. *Architecture versus Content, Sharing of Information*

Hanson defines the "content" of a system to be its small modular features, while its "architecture" is its most important, least modular features. The lesson Lenat took from Eurisko was that architecture is overrated; AIs learn slowly now mainly because they know so little. Thus, AI knowledge needs be to explicitly coded by hand until we have enough to build systems effective at asking questions, reading, and learning for themselves. Prior AI researchers were too comfortable starting every project over from scratch; they needed to join to create larger integrated knowledge bases. This still seems like a reasonable view to Hanson. It also implies that most of the work involved in creating an AI is about gathering knowledge, which could be a gradual process with no single entity taking a lead.

In artificial intelligence in general, young researchers keep coming up with new models, but these generally tend to be variants of the old models, just with new names. The architecture doesn't seem that important there.

Yudkowsky comments that Cyc was supposed to become a powerful AI by accumulating enough knowledge, but so far it doesn't work even as well as Eurisko did. He thinks this is mild evidence against the "content is more important" view. Robin answers that maybe Cyc

just doesn't know enough yet, and that it can do a lot of impressive things already.

Hanson offers the analogy of New York City. Suppose that one said, "New York's a decent city. It's all right. But look at all these architectural failings. Look how this is designed badly or that is designed badly. The roads are in the wrong place or the subways are in the wrong place or the building heights are wrong, the pipe format is wrong. Let's imagine building a whole new city somewhere with the right sort of architecture." Then a new city would be built somewhere else, with a much improved architecture, and people would be invited in. Probably there would not be many comers. For cities architecture does matter, but content is far more important.

Similarly, Hanson thinks that what matters for minds is the content—many things that the mind knows, many routines and strategies—and that there isn't that much at the architectural level that's important. Hanson:

> For similar reasons, I'm skeptical of a blank-slate AI mind-design intelligence explosion. Sure if there were a super mind theory that allowed vast mental efficiency gains all at once, but there isn't. Minds are vast complex structures full of parts that depend intricately on each other, much like the citizens of a city. Minds, like cities, best improve gradually, because you just never know enough to manage a vast redesign of something with such complex inter-dependent adaptations.

Hanson mentions Peter Norvig's recent paper, where Norvig was arguing with Noam Chomsky and saying that it's wrong to expect there

to be a simple elegant theory of linguistics. Instead there are just many messy details that one has to get right, with no key architecture.

Yudkowsky replies that history knows many examples where a single team has gotten far ahead of others. Hanson answers that single teams have gotten far ahead of other teams on narrow issues, while Yudkowsky is postulating an AI that would get far ahead of others on the whole general subject of mental capacity. Yudkowsky's answer is that he is postulating an AI that would get far ahead of others in the narrow, single technology of intelligence. He views intelligence not like a city, which is all over the place, but more like a car, as a machine that takes inputs and has outputs.

Yudkowsky notes that there aren't that many differences between humans and chimps, and that chimps seem to mostly have the same brain areas as humans do. Yudkowsky has an example that he likes to call the "one-wrong-number function": somebody dialing 90% of Yudkowsky's phone number right does not get them a person who is 90% Eliezer Yudkowsky. Likewise, getting 90% of the human architecture correct does not get a being that's 90% capable of human work. The key architectural differences seem to matter a great deal.

Hanson is skeptical about whether it's the architecture that matters the most for humans and chimps, or the lack of social instincts and domesticability for chimps. And although it's true that there was some key change between humans and chimps, that doesn't mean that there'd be a landscape of intelligence where you could make something billions of times faster than humans by just rearranging the architecture.

Yudkowsky notes that for this argument to carry it's not enough to say that content matters. It also needs to be established that there are

no master tricks for learning content faster. The scientific method, for instance, was a master trick that allowed for the faster accumulation of content.

Hanson answers that there's a large literature on economic and ecological innovations, basically saying that the vast majority of innovation consists of small gains. It's lots of little steps over time that slowly make various fields better.

Yudkowsky argues that there's no reason why a single AI couldn't necessarily come up with as many innovations as a community of humans. Although there are six billion people on Earth, that population is not six billion times as smart as a single human. A human brain is four times as large as a chimp's, but four chimps do not equal a single human. Nor could a billion squirrel brains compete with one human brain. Biological brains simply aren't very good at combining their efforts. Buying twice as many scientists doesn't get twice as much science, it gets twice as many science papers.

Making a brain twice as large, with a unified architecture, seems to produce a scaling of output of intelligence that is not even remotely comparable to the effect of taking two brains of fixed size and letting them talk to each other using words. It does not seem at all implausible that an AI that could properly scale to the available computing power could outperform the efforts of six billion people flapping their lips at each other.

9. Modularity of Knowledge

Yudkowsky notes that the abilities we call human are produced in a brain that has a variety of systems specialized for various tasks,

but which work *together* to produce the final abilities. To try to get human-like performance in just one domain is like having a global economy that can only manufacture toasters, not dishwashers or light bulbs. Something like Deep Blue can beat humans in chess in an inhuman way, but to have human-like performance in biology R&D (for example) would require an architecture general enough to also produce human-like performance in other domains. Yudkowsky considers this a fair analogy to the notion that one shouldn't see a global economy that can manufacture toasters but nothing else.

Yudkowsky argues that trading cognitive content between different kinds of AIs is likely to be very hard. In current-day AI, there are few standard databases of preprocessed cognitive content that one can simply buy and plug into an AI system. There are things like databases of stored games of chess, usable with chess-playing programs, but that is not the same as having databases of actual cognitive content.

Even AIs based on the same architecture by the same programmers may be incapable of exchanging information with each other. If two AIs both see an apple for the first time, and they both independently form concepts about that apple, and they both independently build some new cognitive content around those concepts, then their thoughts are effectively written in different languages. By seeing a single apple at the same time, they could identify a concept they both have in mind, and in this way build up a common language—but they would still need a special language designed for sharing knowledge, even if they shared the same source code. With AIs of different architectures, it would be easier to just redo all the cognitive work of learning on one's own, as it is done today. It seems like AIs would have to get very sophisticated before they got over this challenge.

This is also the reason why it's likely to be a single coherent system that undergoes hard takeoff via recursive self-improvement. The same sort of barriers that apply to trading direct cognitive content would also apply to trading changes in cognitive source code. It would be easier for an AI to modify its own source code than to take that modification and sell it to another AI that happens to be written in a different manner. Certain abstract, general insights might be more tradeable, but at that point one is talking about AIs that already understand AI theory, at which point there is likely already a hard takeoff going on.

Suppose that there was a community of diverse AIs which were sophisticated enough to share cognitive content, code changes, and even insights, and there was not yet a hard takeoff. Suppose further that most of the code improvements, algorithmic insights, and cognitive content driving any particular AI were coming from outside that AI—sold or shared—so that the improvements the AI made to itself did not dominate the total velocity. Even in that case, the situation is hard for humans. Even presuming emulations, it will be immensely more difficult to apply any of the algorithmic insights that are tradeable between AIs to the human brain.

Hanson responds that almost all technologies initially come in a vast variety of styles, until they converge to what later seems an obvious configuration. When people begin actually implementing technologies, society figures out the best approaches while network and other scale effects lock in popular approaches. As standards congeal, competitors focus on smaller variations around accepted approaches. Those who stick with odd standards tend to be marginalized. Of course early AI systems take a wide range of incompatible approaches,

but commercial hardware tries a lot harder to match standards and share sources.

Hanson gives the example of automobiles. The people who created the first automobiles merely built a car without worrying about standards. Over time an infrastructure built up, as well as a whole industry involving suppliers, manufacturers, filling stations, repair shops and so on, all of them matched and integrated with each other. In a large real economy of smart machines, there would be standards as well as strong economic pressures to match those standards.

Hanson also mentions programming languages as an example. If a programming language has many users, then compared to a language with a small number of users, the language with a lot of users can accumulate improvements faster. If there is a an AI that is just working on its own, it needs a huge advantage to counter the fact that it is not benefiting from the work of others. In contrast, if different people have different AIs, and everyone who finds a small improvement to their own machine shares it with the others, that community can grow vastly faster than someone trying to do everything on their own. Thus there would again be a pressure to standardize and share.

Hanson says that an effective AI system cannot just be created by building the right architecture and feeding it a lot of raw data; it also needs a considerable amount of content to make sense of it. One could not build an effective cell, or ecosystem, or developed economy, or any other complex system by simply coming up with a good architecture—complex systems require not just good structure, but also lots of good content. Loners who start from scratch rarely beat established groups sharing enough standards to let them share improvements and slowly accumulate content. AI just won't happen

without a whole lot of content. If emulations appear first, perhaps shareable emulation contents could form a different basis for shared improvements.

Yudkowsky suggests that human babies growing up are an example of a good architecture which is then fed large amounts of raw data from the environment. Hanson replies that, in addition to good architecture, a human baby also has large amounts of genetically encoded content about the kind of information to pay attention to, and human babies are also explicitly taught. Yudkowsky says that his visualization of how an AI works would be much like this, only that there would be substantially less genetically coded information at the time of bootup.

10. Local or Global Intelligence Explosion?

Hanson notes that today's economy is highly interdependent—innovations made on one side of the world depend on earlier innovations made on the opposite side of the world. Likewise, raw materials or components of a product may come from a very long distance away. The economy is thus *global*. In contrast, visions of a hard takeoff seem very *local*: technological advances in one small group allow that group to suddenly grow big enough to take over everything. This presumes a very powerful but autonomous area of technology: progress in that area must depend only on advances in the same area. A single group must be able to make great progress in it, and that progress must by itself be sufficient to let the group take over the world. This seems unrealistic, given today's trends.

Yudkowsky notes that there was a brief period when only the USA had nuclear weapons, and they therefore had a decisive military advantage against everyone else. With computing, there was never a moment when one country would have had a decisive advantage over all others. How will things look for AI?

Molecular nanotechnology (MNT) is a hypothetical technology based on the ability to build structures to complex, atomic specifications. In theory, sufficiently advanced MNT would allow one to construct things on the atomic level, reconfiguring local matter to work as the raw material for whatever was being produced.

Yudkowsky discusses the impact of MNT on the local/global question. In theory, MNT would allow one to create a purely local manufacturing complex, producing all the materials on one site. With the ability to produce solar cells, the factory could also obtain its own energy. As MNT theoretically allows the creation of self-replicating machines, it may be enough to merely build the initial machine, and it will build more.

A research group developing better software is still reliant on outside groups for hardware and electricity. As long as this is the case, they cannot use their innovations to improve their hardware or to drive down the cost of electricity—at least not without giving that knowledge to outside groups. Any innovational cascades will then only affect a part of what makes the group productive, setting an upper limit on the extent to which innovations can help the group. The more capabilities are localized into one place, the less people will depend on their trade partners, the more they can cascade locally (apply their improvements to yield further improvements), and the more a "critical cascade"/FOOM sounds plausible.

Analysis. Hall's paper "Engineering Utopia"[15] makes essentially this argument, noting that AIs would still benefit from trading with the rest of the world, but that at some point it would become possible for superfast AIs to trade exclusively among themselves, at which point their speed of development would FOOM far past humanity's.

There's an analogy to Amdahl's law here.[16] The law states that if a fraction f of a program's performance can be parallelized, then the speedup given by n processors instead of one is $1/[(1-f)+(f/n)]$. More generally, if a fraction f of a group's performance depends on a specific capability, then the overall performance improvement given by improving that capability by a factor of n is proportional to $1/[(1-f)+(f/n)]$.

On the other hand, MNT is a very advanced technology. Yudkowsky notes that current-day work on nanotechnology is still very global, and it would not be inconceivable that this trend would continue even as MNT improved, due to the normal benefits of specialization and division of labor. Several countries might race toward better and better MNT, none of them achieving a decisive advantage. MNT is not necessarily a sudden discontinuity by itself: it might allow a relatively smooth trajectory.

However, a discontinuity is likely to happen if emulations are developed after MNT. Nanocomputers are very powerful, and the first emulations might be able to run a thousand times faster than biological humans the moment they became viable enough to do scientific research. Even if one country only had a one-day advantage compared to all the others, the thousandfold speed advantage would rapidly accumulate. In just an hour of time, the emulations could do

a year's worth of research. This might allow them to develop and implement further technologies which allowed them to run even faster.

If emulations were gradually developed at a time when computers were too slow to run them quickly, things would be different. The first high-fidelity emulations, running at a hundredth of human speed, would grant no special advantage.

Yudkowsky says that his main purpose with this discussion is to illustrate the point that, as optimizers become more self-swallowing, races between them become more unstable. The less dependent something is on outside forces, the stronger the effect of innovation cascades on its capabilities. If everything could be built practically instantly via MNT, and research could be conducted by emulations running at far higher speeds, then a single theoretical breakthrough could precipitate an instant international military crisis. The situation would be quite different from today, where there is a long delay between discovery and implementation, and most discoveries never amount to anything.

Hanson notes that there is no law of increasingly local production. The locality of manufacturing comes from tradeoffs between economies and diseconomies of scale. Things can often be made cheaper in big centralized plants, especially if located near key inputs. When processing bulk materials, for example, there is a rough two-thirds-cost power law: throughput goes as volume, while the cost to make and manage machinery tends to go as surface area. But it costs more to transport products from a few big plants. Local plants can offer more varied products, explore more varied methods, and deliver cheaper and faster.

Innovation and adaption to changing conditions can be faster or slower at centralized plants, depending on other details. Politics sometimes pushes for local production to avoid dependence on foreigners, and at other times pushes for central production to make succession more difficult. Smaller plants can better avoid regulation, while larger ones can gain more government subsidies. When formal intellectual property is weak (the usual case), producers may prefer to make and sell parts instead of selling recipes for making parts. Even in an MNT-dominated economy, production may still be global due to the same economic reasons as it is today.

Yudkowsky replies that he has no objections to most of this, but one can serve quite a lot of needs by having "nanoblocks" that reconfigure themselves in response to demands. He thinks that this would be a localizing force with respect to production, and a globalizing force with respect to design.

Hanson replies that if, as Yudkowsky accepts, manufacturing may not be very local, then it would be harder for an AI to build the physical equipment that's needed for taking over the world undetected. Yudkowsky's response is that an intelligent-enough AI might very well come up with the needed plausible cover stories and, for example, buy mail-order proteins undetected. Hanson responds that taking over the world might require more than a few mail order proteins, to which Yudkowsky responds that it might not—ribosomes are reasonably general molecular factories and quite capable of self-replication.

Hanson says that he is just highlighting the extreme degree of intelligence postulated. The hypothetical AI, which has made no visible outside mark beyond mail-ordering a few proteins, knows enough to

use those proteins to build a physically small manufacturing industry that is more powerful than the entire rest of the world.

11. Wrap-up

In the end, Yudkowsky and Hanson fail to reach agreement.

Hanson summarizes Yudkowsky's view: "I read Eliezer as fearing that developers, insurers, regulators, and judges, will vastly underestimate how dangerous are newly developed AIs. Eliezer guesses that within a few weeks a single AI could grow via largely internal means from weak and unnoticed to so strong it takes over the world, with no weak but visible moment between when others might just nuke it. Since its growth needs little from the rest of the world, and since its resulting power is so vast, only its values would make it treat others as much more than raw materials. But its values as seen when weak say little about its values when strong. Thus Eliezer sees little choice but to try to design a theoretically clean AI architecture allowing near-provably predictable values when strong, to in addition design a set of robust good values, and then to get AI developers to adopt this architecture/values combination."

Hanson notes that he finds Yudkowsky's suggestions of rapid growth unpersuasive: normally dozens of relevant factors are co-evolving, some of them feeding circularly into each other. Yet it usually all adds up to exponential growth, with rare jumps to faster growth rates.

Hanson thinks that locality is the point of greatest disagreement. He asks us to imagine a scenario with a large community of AI developers selling AI to customers, in which AIs got mostly better by ac-

cumulating better content, and the rate of accumulation mainly depended on previous content. In this scenario the AI section of the economy might grow pretty quickly, but it would be hard to imagine one AI project zooming vastly ahead of others. AI architecture would have relatively little significance.

So the disagreement may be a disagreement about how powerful architecture is in AI, and how many architectural insights could be found in a given time. If there were a series of twenty deep, powerful insights, each of which made a system twice as effective—just enough to let it find the next insight—it would add up to a factor of one million. But this still wouldn't be enough to let a single AI take over the world.

Hanson: "This scenario seems quite flattering to Einstein-wannabes, making deep-insight-producing Einsteins vastly more valuable than they have ever been, even in percentage terms. But when I've looked at AI research I just haven't seen it. I've seen innumerable permutations on a few recycled architectural concepts, and way too much energy wasted on architectures in systems starved for content, content that academic researchers have little incentive to pursue. So we have come to: What evidence is there for a dense sequence of powerful architectural AI insights? Is there any evidence that natural selection stumbled across such things?"

Yudkowsky notes that if, as Hanson predicts, the AI section of the economy might grow rapidly but without much chance for one AI project to zoom ahead of the others, the AIs as a group might still zoom ahead of the humans. It could then be a huge benefit to all AIs to simply eliminate the "statue-slow, defenseless, noncontributing humans."

Hanson's response is that coordination is hard, and humans have built a great number of institutions for the sake of aiding coordination. Since coordination depends crucially on institutions, AIs would need to preserve those institutions as well. So AIs would not want to threaten the institutions they use to keep the peace among themselves. It is far from easy to coordinate to exterminate humans while preserving such institutions.

Yudkowsky disagrees, believing that much of today's cooperation comes rather from humans having a sense of honor and an internalized group morality, rather than from a rational calculation to avoid conflict in order to maximize resources: "If human beings were really genuinely selfish, the economy would fall apart or at least have to spend vastly greater resources policing itself—think Zimbabwe and other failed states where police routinely stop buses to collect bribes from all passengers, but without the sense of restraint: the police just shoot you and loot your corpse unless they expect to be able to extract further bribes from you in particular." We thus cannot depend on AIs maintaining and using our cooperation-preserving institutions in such a way that would protect human interests.

Hanson replies that such a position not only disagrees with his opinions on the sources and solutions of coordination problems, it also disagrees with the opinions of most economists. He admits that genuinely selfish humans would have to spend more resources to coordinate with those that they were in daily contact with, because we have evolved adaptations which increase our ability to coordinate on a small scale. But we do not have such adaptations for large-scale coordination, and have therefore created institutions to carry out that task. Large-scale coordination in society of selfish humans would be

just as easy, and since such coordination depended crucially on institutions, AIs would need to preserve those institutions as well.

Yudkowsky summarizes his view of the debate in turn. His biggest disagreement is over the way that Hanson frames his analyses: "It's that there are a lot of opaque agents running around, little black boxes assumed to be similar to humans, but there are more of them and they're less expensive to build/teach/run. They aren't even any faster, let alone smarter. (I don't think that standard economics says that doubling the population halves the doubling time, so it matters whether you're making more minds or faster ones.) . . . So that world looks like this one, except that the cost of 'human capital' and labor is dropping according to (exogenous) Moore's Law, and it ends up that economic growth doubles every month instead of every sixteen years—but that's it."

Yudkowsky admits that Hanson has a strong point about "unvetted abstractions," but thinks that there's something wrong with using it as justification for defending the superiority of models that are made up of many human-like black boxes whose fundamental behavior is never altered. He points out that his own simple model of Moore's Law, which predicted a vastly faster speed of development once the people who developed computers were themselves running on computers, was probably as well-vetted as Hanson's earlier paper on economic growth given machine intelligence.[17] Both are models of a sort that haven't been used before, in domains not actually observed, and both predict a future quite different from the world we see. Yudkowsky suspects that Hanson is actually finding Yudkowsky's conclusions objectionable for other reasons, and that Hanson thus imposes a stricter burden of proof on the kinds of abstractions that Yudkowsky

uses than the ones Hanson himself uses, without properly explaining why.

Hanson answers that a community of thousands of specialists has developed over decades examining models of total system growth. He has not just talked about vetting, but also offered more detailed reasons of why Yudkowsky's model seems unsatisfactory.

Yudkowsky has no problem with the specific reasons Hanson offers, it's just the "insufficiently vetted" part of the argument that he finds difficult to engage with, as it doesn't let him know the exact criteria by which the models are being judged. Without such criteria, it seems like an appeal to authority, and while Yudkowsky says that he does not reject authority in general, the models of the economists are all entirely tested on the behavior of humans. It is hard for him to believe that economists have taken into account the considerations involved in translating the special case of humans into a more general model, when several basic assumptions may be broken. He expects the economists' models to only work for describing humans.

Yudkowsky also says that he sees his view of an AI possibly going from relatively limited intelligence to superintelligence in less than a week as an "antiprediction"—a prediction that sounds very startling, but actually isn't. He gives the example of a man who was asked what he thought of his odds of winning the lottery, and who replied "fifty–fifty—either I win or I don't." Only a small number of all the possible combinations of lottery balls will allow a person to win, so the most probable prediction is that the man won't win. One may be tempted to object to such a prediction, saying that the other person doesn't have enough evidence for it, but in reality they are making a mistake by focusing excessively on such a low-probability event in the first place.

Likewise, "less than a week" may sound fast in human terms. But a week is 10^{49} Planck intervals, and if one looks at the various timescales during which different events occur—from Planck intervals to the age of the universe—then it seems like there's nothing special about the timescale that humans happen to live on. An AI running on a 2 GHz processor could perform 10^{15} serial operations in a week, and 10^{19} serial operations in a century. If an AI is likely to improve itself to superintelligence in the first place, then it is likely to do it in less than 10^{15} or more than 10^{19} serial operations, since the region between them isn't all that wide of a target. So it will take less than a week or more than a century, in which case any faster AI will beat the slower one.

Hanson finds this unpersuasive and feels that the core questions involve the relative contribution of architecture and content in minds, as well as how easy it will be to quickly find a larger number of powerful architectural improvements. Yudkowsky thinks that the existence of visible flaws in human cognition implies a lack of diminishing returns near the human level, as one can go past the human level by simply correcting the flaws. Hanson disagrees, as simply being aware of the flaws doesn't imply that they're easy to correct.

* * *

1. Irving John Good, "Speculations Concerning the First Ultraintelligent Machine," in *Advances in Computers*, ed. Franz L. Alt and Morris Rubinoff, vol. 6 (New York: Academic Press, 1965), 31–88, doi:10.1016/S0065-2458(08)60418-0; Eliezer Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan M. Ćirković (New York: Oxford University Press, 2008), 308–345; David John Chalmers, "The Singularity: A Philosophical

Analysis," *Journal of Consciousness Studies* 17, nos. 9–10 (2010): 7–65, http://www.ingentaconnect.com/content/imp/jcs/2010/00000017/f0020009/art00001; Luke Muehlhauser and Anna Salamon, "Intelligence Explosion: Evidence and Import," in *Singularity Hypotheses: A Scientific and Philosophical Assessment*, ed. Amnon Eden et al., The Frontiers Collection (Berlin: Springer, 2012).

2. Nick Bostrom, "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards," *Journal of Evolution and Technology* 9 (2002), http://www.jetpress.org/volume9/risks.html; Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk"; Muehlhauser and Salamon, "Intelligence Explosion."

3. Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk"; Chalmers, "The Singularity"; Luke Muehlhauser and Louie Helm, "The Singularity and Machine Ethics," in Eden et al., *Singularity Hypotheses*.

4. Eliezer Yudkowsky, *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*, The Singularity Institute, San Francisco, CA, June 15, 2001, http://intelligence.org/files/CFAI.pdf; Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk"; Eliezer Yudkowsky, *Complex Value Systems are Required to Realize Valuable Futures* (The Singularity Institute, San Francisco, CA, 2011), http://intelligence.org/files/ComplexValues.pdf; Nick Bostrom and Eliezer Yudkowsky, "The Ethics of Artificial Intelligence," in *Cambridge Handbook of Artificial Intelligence*, ed. Keith Frankish and William Ramsey (New York: Cambridge University Press, forthcoming).

5. Hanson, "If Uploads Come First"; Hanson, "Economic Growth Given Machine Intelligence"; Hanson, "Economics of the Singularity"; Robin Hanson, "Meet the New Conflict, Same as the Old Conflict," *Journal of Consciousness Studies* 19, nos. 1–2 (2012): 119–125, http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00008.

6. Shane Legg and Marcus Hutter, "Universal Intelligence: A Definition of Machine Intelligence," *Minds and Machines* 17, no. 4 (2007): 391–444, doi:10.1007/s11023-007-9079-x.

7. Hanson, "Economic Growth Given Machine Intelligence."

8. Nicola Jones, "Middle-eastern Farmers 'Civilised' Europe," *New Scientist*, August 5, 2002, accessed June 26, 2013, http : / / www . newscientist . com / article / dn2634 - middleeastern - farmers - civilised - europe.html.

9. Laura Spinney, "The Gene Chronicles," *New Scientist*, February 7, 2004, no. 2433, accessed June 26, 2013, http://www.newscientist.com/article/mg18124335.200.

10. Angus Maddison, "Measuring and Interpreting World Economic Performance 1500–2001," *Review of Income and Wealth* 51, no. 1 (2005): 1–35.

11. *2007 World Population Datasheet* (Washington, DC: Population Reference Bureau, August 2007), accessed June 26, 2013, http://www.prb.org/pdf07/07WPDS_Eng.pdf.

12. Robin Hanson, "Natural Genocide," *Overcoming Bias* (blog), June 18, 2008, http : / / www.overcomingbias.com/2008/06/natural - genocid.html.

13. Robin Hanson, "Britain Was Too Small," *Overcoming Bias* (blog), June 19, 2008, http: //www.overcomingbias.com/2008/06/britain - was - too.html.

14. Hanson, "Dreams of Autarky."

15. John Storrs Hall, "Engineering Utopia," in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, ed. Pei Wang, Ben Goertzel, and Stan Franklin, Frontiers in Artificial Intelligence and Applications 171 (Amsterdam: IOS, 2008), 460–467.

16. Gene M. Amdahl, "Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities," in *Proceedings of the April 18–20, 1967, Spring Joint Computer Conference—AFIPS '67 (Spring)* (New York: ACM Press, 1967), 483–485, doi:10.1145/1465482.1465560.

17. Hanson, "Economic Growth Given Machine Intelligence."

62

Intelligence Explosion Microeconomics

Eliezer Yudkowsky

6 May 2013

Editor's Note: This chapter was originally published as a technical report by the Machine Intelligence Research Institute. The latest version of this report can be found at http://intelligence.org/files/IEM.pdf.

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 correctly notes,[1] 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, 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.

1. The Intelligence Explosion: Growth Rates of Cognitive Reinvestment

In 1965, I. J. Good[2] published a paper titled "Speculations Concerning the First Ultraintelligent Machine" 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.[3]

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" for a compact overview.[4] See also David Chalmers's 2010 paper,[5] the responses, and Chalmers's reply.[6]

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 "Economic Growth Given Machine Intelligence" considers emulations of scanned human brains (a.k.a. *ems*):[7] 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[8] are driving scientific and technological progress is difficult enough that previous attempts at formal 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 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.[9]

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[10]—see again Chalmers[11] or Muehlhauser and Salamon.[12] 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 put it:

> The key issue is the "proportionality thesis" saying that among systems of certain class, an increase of δ in intelligence will yield an increase of δ in the intelligence of systems that these systems can design.[13]

To illustrate the core question, let us consider a nuclear pile undergoing a fission reaction.[14] 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 emit-

ted 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 fifty-six and fifty-seven of uranium bricks, but as layer fifty-seven was added, wooden rods covered with neutron-absorbing 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."[15] 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 ex-

actly 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.[16]

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 predicts world economic doubling times of be-

tween a month and a year.[17] 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*.[18] All compound interest returned on discoveries such as the invention 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."[19]

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 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,[20] 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 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.[21]

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 order-

ing. 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.[22]

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.

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 ran-

domized 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 in which "self-improvement" requires an increase in Kolmogorov complexity of an isolated system,[23] or Bringsjord's definition in which a Turing machine is only said to self-improve if it can raise itself into a class of hypercomputers.[24] 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.[25]

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.[26]

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 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 be-

fore then, and numerous other issues that don't seem to interact very strongly with the returns on cognitive investment *qua* cognitive investment.[27]

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 and Bostrom point out, most possible preferences imply capability increase as an instrumental motive.[28] 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,[29] and there will always be *some* special case of an agent that chooses not to carry out any given deed.[30] 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. Policies are motivated by differentials of expected util-

ity; 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.[31] 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[32] so as to arrive at an output which corresponds intuitively to the AI "doing the right thing." See also "Artificial Intelligence as a Positive and Negative Factor in Global Risk."[33]

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 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[34] and Muehlhauser and Salamon.[35] 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 an-

swering the epistemic question of which actions would result in great numbers of paperclips, then it will be a very instrumentally powerful agent.[36]

The *Complexity of Value Thesis* says that human values are complex in the sense of having high algorithmic (Kolmogorov) complexity.[37] Even idealized forms of human value, such as reflective equilibrium[38] or ideal advisor theories[39]—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., have no privileged reason to be readily reducible to each other or to anything else.[40] The Complexity of Value Thesis is that to realize valuable outcomes, an AI must have complex information in its utility function; it also will not suffice to tell it to "just make humans happy" or any other simplified, compressed principle.[41]

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.[42] 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 Instrumen-

tal 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 intelligence reinvested into AI will only get logarithmic re-

turns 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.[43]

- 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 sapi-*

ens having roughly four times as much cortical volume and six times as much prefrontal cortex.[44]

- 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.[45]

- 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 questioning Moore's Law's validity;[46] 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.[47]

- 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*.[48]

- 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.[49] 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[50]) 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.[51]) A true cynic might claim that, in many fields, exponentially greater investment in science is yielding a roughly constant amount of annual progress—sublogarithmic returns![52]

- This graph 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.[53]

- 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 speedups like the invention of agriculture.[54] Or obviously it's analogous to other foundational changes in the production of complex order, such as human intelligence or self-replicating life.[55] 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."[56] 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 sam-

pled 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.[57]

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.[58] 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?

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, Kurzweil$_{2001}$ regarded Moore's Law as the primary phenomenon and considered machine superintelligence a secondary phenomenon which ought to assume whatever shape was required to keep the primary phenomenon on track.[59]

You could even imagine arguing (though 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,[60] 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.)

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.[61] 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.

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.[62] (See, e.g., the model of Moravec.[63])

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 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.[64] 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*.

The "outside view" is a term from the heuristics and biases program in experimental psychology.[65] 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*[66] 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:

> 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.[67]

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 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.[68]

More on this analysis can be found in Hanson's "Long-Term Growth as a Sequence of Exponential Modes."[69]

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 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.[70]

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[71]).

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,[72] 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.[73] 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.[74]

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—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.[75] 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!' as Conversation-Halter."[76] 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 try-

ing 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.[77] In very concrete terms—leaving out all the abstract principles, microfoundations, 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,[78] 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,[79] or days or seconds.[80] 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-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.[81]
- An AI can continue reinvesting these gains until it has a huge cognitive problem-solving advantage over humans.
- 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.[82] 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.[83] 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 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.[84] 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.[85] 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 rea-

son 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.[86] 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.[87]

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 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[88] *or* strongly diminishing cognitive returns on adding further neurons,[89] then we should expect the 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.[90] 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.[91]

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[92] 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).[93] 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 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.[94]

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 im-

provements 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.[95] 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.
- 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.[96]

"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.

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,[97] 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.[98] Such humans could, in

principle, if immortal 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.[99] 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 devel-

oped 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 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 sup-

pose 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.[100]

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.[101] What separates the early twenty-first century from the eleventh?

Gregory Clark 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.[102]

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,[103] 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.[104] 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).[105]

Similarly, the "hundred-step rule" in neuroscience 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.[106] 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 2×10^{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.[107] 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?"[108] 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.

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.[109] 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,[110] and (b) that fast thinking may prove useless without fast sensors and fast motor manipulators.[111]

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, *ce-*

teris 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 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.[112]

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 op-

eration 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,[113] 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.[114] 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.[115] 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.[116]

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 years of "time,"[117] 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 prenanotechnological 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.

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, say-

ing that he had never expended so much effort in his life, and later wrote a book about the game,[118] 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.[119]

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 re-

construct 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 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,"[120] 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.[121] 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 lim-

ited 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 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[122] 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 them-

selves, 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.[123] But in models with a localized FOOM 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.

By the time we are done observing that not all humans are scientists, that not all scientists are productive, that not all productive sci-

entists 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.[124]

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* 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,[125] 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.[126] 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.[127]

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.[128] 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 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.[129] 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 C depends on both A and B, both A and B must be highly prevalent before there is significant pressure favoring C.[130] 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 hundredfold* among humans since the start of agriculture.[131] 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 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 mutation to account for apparently superlinear historical progress.[132] 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 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.[133] 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.[134]

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 backchaining, and learned abstraction would carry steadily onward—rather than the other way around. Difficulty curves for intelligent engineers 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).[135] 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 hu-

man 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* 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.[136]

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 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 analyzed this question and its several possible answers given the present state of controversy regarding how to reason about

anthropic probabilities.[137] 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 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.[138]

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.[139]

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, 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.[140] 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."[141] But the advice from anthropics is that the road from mice to humans 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.[142] 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 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[143] technology-driven productivity growth[144]—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.[145] 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 caused hominids to pull far ahead of other primates, if hominid improvements had been spreading to primates via horizontal gene transmission.[146]

Thus we can sketch two widely different possible scenarios for an intelligence explosion, at opposite extremes along multiple dimensions, as follows:[147]

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 of any competitors it wishes to fetter. The final result is a Bostrom-style "singleton."[148]

Extremely global takeoff:

- The emergence of good, successful machine intelligence techniques greatly winnows the plethora of visionary prototypes we see nowadays.[149] 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 rapacious hardscrapple frontier folk.[150])

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,[151] then it's implausible that the next runner-up will happen to find it in the same five-second window.[152] 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.[153])

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 equiva-

lent 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.[154]

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."[155] In the highly global scenario we get incremental improvements by having only some AIs be human-Friendly,[156] 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" competitors to later become a singleton.[157] 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.[158] 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.[159] 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.[160]) 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 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.[161]

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, 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 op-

portunities 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.[162] I am tempted to say at this point, "Thankfully, that is 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, 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.[163] 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.[164] 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.

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).[165]

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.[166]
- Over this model space, weight prior probability by simplicity and regularity.
- 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.[167] 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.[168] 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 book *Mind Children*[169] 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 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.[170] 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 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.[171] 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 answer 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:

- 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?[172]

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, 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.[173] (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. 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).

* * *

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

2. 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.

3. Good, "Speculations Concerning the First Ultraintelligent Machine."

4. Muehlhauser and Salamon, "Intelligence Explosion."

5. Chalmers, "The Singularity."

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

7. Hanson, "Economic Growth Given Machine Intelligence."

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

9. Sandberg, "An Overview of Models of Technological Singularity."

10. A.k.a. general AI, a.k.a. strong AI, a.k.a. Artificial General Intelligence. See Cassio Pennachin and Ben Goertzel, "Contemporary Approaches to Artificial General Intelligence," in Goertzel and Pennachin, *Artificial General Intelligence*, 1–30.

11. Chalmers, "The Singularity."

12. Muehlhauser and Salamon, "Intelligence Explosion."

13. Chalmers, "The Singularity."

14. 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.

15. Rhodes, *The Making of the Atomic Bomb*.

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

17. Hanson, "Long-Term Growth as a Sequence of Exponential Modes."

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

19. 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, "Artificial Intelligence as a Positive and Negative Factor in Global Risk").

    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.

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

21. 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*?"

22. "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.

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

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

25. 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.

26. 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.

27. 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.

28. Stephen M. Omohundro, "The Basic AI Drives," in Wang, Goertzel, and Franklin, *Artificial General Intelligence 2008*, 483–492; Nick Bostrom, "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents," in "Theory and Philosophy of AI," ed. Vincent C. Müller, special issue, *Minds and Machines* 22, no. 2 (2012): 71–85, doi:10.1007/s11023-012-9281-3.

29. Muehlhauser and Salamon, "Intelligence Explosion."

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

31. 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 *Neuromancer*.[174] See also the sort of assumptions used in Robert Freitas's *Some Limits to Global Ecophagy*,[175] 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.

32. Luke Muehlhauser and Chris Williamson, "Ideal Advisor Theories and Personal CEV" (2013), http://intelligence.org/files/IdealAdvisorTheories.pdf.

33. Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk."

34. Armstrong, "General Purpose Intelligence."

35. Muehlhauser and Salamon, "Intelligence Explosion."

36. 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.

37. Eliezer Yudkowsky, "Complex Value Systems in Friendly AI," in *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings*, Lecture Notes in Computer Science 6830 (Berlin: Springer, 2011), 388–393, doi:10.1007/978-3-642-22887-2_48; Muehlhauser and Helm, "The Singularity and Machine Ethics."

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

39. Connie S. Rosati, "Persons, Perspectives, and Full Information Accounts of the Good," *Ethics* 105, no. 2 (1995): 296–325, doi:10.1086/293702.

40. See William K. Frankena, *Ethics*, 2nd ed., Foundations of Philosophy Series (Englewood Cliffs, NJ: Prentice-Hall, 1973), chap. 5 for one list of commonly stated terminal values.

41. 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.

42. Omohundro, "The Basic AI Drives"; Bostrom, "The Superintelligent Will."

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

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

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

46. Ilkka Tuomi, "The Lives and the Death of Moore's Law," *First Monday* 7, no. 11 (2002), http://firstmonday.org/ojs/index.php/fm/article/view/1000/921.

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

48. 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.

49. Hanson, "Long-Term Growth as a Sequence of Exponential Modes."

50. National Science Board, *Science and Engineering Indicators 2012*, NSB 12-01 (Arlington, VA, 2012), chap. 5, http://www.nsf.gov/statistics/seind12/start.htm.

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

52. 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.

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

54. Hanson, "Outside View of the Singularity."

55. Eliezer Yudkowsky, "Optimization and the Singularity," *Less Wrong* (blog), June 23, 2008, http://lesswrong.com/lw/rk/optimization_and_the_singularity/; Eliezer Yudkowsky, "Surprised by Brains," *Less Wrong* (blog), November 23, 2008, http://lesswrong.com/lw/w4/surprised_by_brains/; Eliezer Yudkowsky, "The First World Takeover," *Less Wrong* (blog), November 19, 2008, http://lesswrong.com/lw/w0/the_first_world_takeover/.

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

57. 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.

58. 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*,[176] but Wikipedia's cited timelines are shorter term than I remember.

59. 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 suggests that Kurzweil has changed his position since then.[177]

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

61. 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.

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

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

64. 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.)

65. Daniel Kahneman and Dan Lovallo, "Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking," *Management Science* 39, no. 1 (1993): 17–31, doi:10.1287/mnsc.39.1.17.

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

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

68. Hanson, "Outside View of the Singularity."

69. Hanson, "Long-Term Growth as a Sequence of Exponential Modes."

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

71. Hanson, "Long-Term Growth as a Sequence of Exponential Modes."

72. Michael A. McDaniel, "Big-Brained People are Smarter: A Meta-Analysis of the Relationship between In Vivo Brain Volume and Intelligence," *Intelligence* 33, no. 4 (2005): 337–346, doi:10.1016/j.intell.2004.11.005.

73. 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.

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

75. 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.

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

77. 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 revolved entirely around equations consisting of uppercase 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.

78. 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.

79. 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.

80. 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.

81. 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.)

82. "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.

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

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

85. 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.

86. 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.

87. 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, "Belief in the Singularity is Logically Brittle" for the appropriate citation by a distinguished scientist.

    We can be reasonably extremely confident that human intelligence does not take advantage of quantum computation.[178] The computing elements of the brain are too large and too hot.

88. 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.

89. 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.

90. Marcia S. Ponce de León et al., "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, no. 37 (2008): 13764–13768, doi:10.1073/pnas.0803917105

    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.

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

92. 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."[179]

93. 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.

94. 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.

95. 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."

96. 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.

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

98. 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.

99. Note the resemblance to the standard reply to Searle's Chinese Room argument.[180]

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

101. See again section 3.11.

102. Clark, *A Farewell to Alms*.

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

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

105. 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.)

106. Feldman and Ballard, "Connectionist Models and Their Properties."

107. 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.

108. I was then seventeen years old.

109. 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.

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

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

112. Andrew Wiles, "Modular Elliptic Curves and Fermat's Last Theorem," *Annals of Mathematics* 142, no. 3 (1995): 443–551, doi:10.2307/2118559.

113. 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.

114. 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.

115. 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.

116. 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!"

117. Albeit, in accordance with the general theme of embarrassingly overwhelming human inefficiency, the actual thought processes separating Yudkowsky_{1997} from 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.

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

119. 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.

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

121. 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.

122. 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.

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

124. 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[181] or the Structure Mapping Engine[182] 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 that this is not a good model.[183] So not quite *that* low.

125. Terrence Deacon's *The Symbolic Species*[184] 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.

126. William H. Calvin, *A Brief History of the Mind: From Apes to Intellect and Beyond* (New York: Oxford University Press, 2004), chap. 5.

127. 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?

128. For an explanation of "protolanguage" see Derek Bickerton, *Adam's Tongue: How Humans Made Language, How Language Made Humans* (New York: Hill & Wang, 2009).

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

130. 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.

131. John Hawks et al., "Recent Acceleration of Human Adaptive Evolution," *Proceedings of the National Academy of Sciences of the United States of America* 104, no. 52 (2007): 20753–20758, doi:10.1073/pnas.0707650104.

132. 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.

133. Williams, *Adaptation and Natural Selection*.

134. 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.

135. 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*.

136. 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*.

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

138. Hanson, "Must Early Life Be Easy? The Rhythm of Major Evolutionary Transitions."

139. 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.

140. 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.

141. 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 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.

142. 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.

143. 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.[185] This is not *fast*, but it is exponential.

144. 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.)

145. 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.

146. 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.

147. 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.

148. Bostrom, "What is a Singleton?"

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

150. Hanson, "The Rapacious Hardscrapple Frontier."

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

152. 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.

153. 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.

154. 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 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.

155. 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.

156. 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.

157. 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.

158. Humans count as human-equivalent intelligences.

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

160. Martin Pelikan, David E. Goldberg, and Erick Cantú-Paz, "Linkage Problem, Distribution Estimation, and Bayesian Networks," *Evolutionary Computation* 8, no. 3 (2000): 311–340, doi:10.1162/106365600750078808.

161. "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.)

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

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

164. 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.

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

166. 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.

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

168. Yes, that is a joke.

169. Moravec, *Mind Children*.

170. See also *The Moon is a Harsh Mistress*[186] 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*, compared this to the ancient theory that dirty shirts and straw would spontaneously generate mice.[187]

171. 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.

172. 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.

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

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

175. Freitas, "Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations."

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

177. Sandberg, "An Overview of Models of Technological Singularity."

178. Max Tegmark, "Importance of Quantum Decoherence in Brain Processes," *Physical Review E* 61, no. 4 (2000): 4194–4206, doi:10.1103/PhysRevE.61.4194.

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

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

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

182. Brian Falkenhainer and Kenneth D. Forbus, "The Structure-Mapping Engine: Algorithm and Examples," *Artificial Intelligence* 41, no. 1 (1990): 1–63, doi:10.1016/0004-3702(89)90077-5.

183. David John Chalmers, Robert M. French, and Douglas R. Hofstadter, "High-Level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology," *Journal of Experimental and Theoretical Artificial Intelligence* 4, no. 3 (1992): 185–211, doi:10.1080/09528139208953747.

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

185. Cowen, *The Great Stagnation*.

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

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

Bibliography

2007 World Population Datasheet. Washington, DC: Population Reference Bureau, August 2007. Accessed June 26, 2013. http://www.prb.org/pdf07/07WPDS_Eng.pdf.

Alcor Life Extension Foundation. "Alcor Membership Statistics." April 30, 2013. Accessed July 28, 2013. http://www.alcor.org/AboutAlcor/membershipstats.html.

———. "Frequently Asked Questions." Accessed July 28, 2013. http://www.alcor.org/FAQs/index.html.

———. "Scientists' Cryonics FAQ." Accessed July 28, 2013. http://www.alcor.org/sciencefaq.htm.

Amdahl, Gene M. "Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities." In *Proceedings of the April 18–20, 1967, Spring Joint Computer Conference—AFIPS '67 (Spring)*, 483–485. New York: ACM Press, 1967. doi:10.1145/1465482.1465560.

Armstrong, Stuart. "General Purpose Intelligence: Arguing the Orthogonality Thesis." *Analysis and Metaphysics* (forthcoming). Preprint at http : / / lesswrong . com / lw / cej / general _ purpose _ intelligence _ arguing_the/.

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

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

Benford, Gregory, Alaxander Bolonkin, Nick Bostrom, Kevin Q. Brown, Manfred Clynes, L. Stephen Coles, Daniel Crevier, et al. "Scientists' Open Letter on Cryonics." Accessed July 24, 2013. http : / / www . evidencebasedcryonics.org/scientists-open-letter-on-cryonics/.

Best, Ben. "Cryonics — Frequently Asked Questions (FAQ)." 2004. Last revised August 22, 2012. http : / / www . benbest . com / cryonics / CryoFAQ.html.

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

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

Bostrom, Nick. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9 (2002). http://www.jetpress.org/volume9/risks.html.

———. "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, no. 2 (2012): 71–85. doi:10.1007/s11023-012-9281-3.

———. "What is a Singleton?" *Linguistic and Philosophical Investigations* 5, no. 2 (2006): 48–54.

Bostrom, Nick, and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence." In *Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William Ramsey. New York: Cambridge University Press, forthcoming.

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

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

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

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

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

Chalmers, David John, Robert M. French, and Douglas R. Hofstadter. "High-Level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology." *Journal of Experimental and Theoretical Artificial Intelligence* 4, no. 3 (1992): 185–211. doi:10.1080/09528139208953747.

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

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

Copeland, Michael V. "How to Make Your Business Plan the Perfect Pitch." *Business 2.0*, September 1, 2005. http://money.cnn.com/magazines/business2/business2_archive/2005/09/01/8356496/.

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

Darwin, Michael G., Chana de Wolf, and Aschwin de Wolf. "Is That What Love Is? The Hostile Wife Phenomenon in Cryonics." *Evidence Based Cryonics* (blog), 2008. http://www.evidencebasedcryonics.org/is-that-what-love-is-the-hostile-wife-phenomenon-in-cryonics/.

Dawes, Robyn M. *Rational Choice in An Uncertain World*. 1st ed. Edited by Jerome Kagan. San Diego, CA: Harcourt Brace Jovanovich, 1988.

De Mesquita, Bruce Bueno, Alastair Smith, Randolph M. Siverson, and James D. Morrow. *The Logic of Political Survival*. Cambridge, MA: MIT Press, 2003.

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

Douglas, Richard W., Jr. "Site Value Taxation and Manvel's Land Value Estimates." *American Journal of Economics and Sociology* 37, no. 2 (1978): 217–223. http://www.jstor.org/stable/3486442.

Drexler, K. Eric. *Engines of Creation*. Garden City, NY: Anchor, 1986.

The Economist. "House of Cards." May 29, 2003. http://www.economist.com/node/1794873.

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

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

Engelbart, Douglas C. *Augmenting Human Intellect: A Conceptual Framework*. Technical report. Menlo Park, CA: Stanford Research Institute, October 1962. http://www.dougengelbart.org/pubs/augment-3906.html.

Falkenhainer, Brian, and Kenneth D. Forbus. "The Structure-Mapping Engine: Algorithm and Examples." *Artificial Intelligence* 41, no. 1 (1990): 1–63. doi:10.1016/0004-3702(89)90077-5.

Feldman, J. A., and Dana H. Ballard. "Connectionist Models and Their Properties." *Cognitive Science* 6, no. 3 (1982): 205–254. doi:10.1207/s15516709cog0603_1.

Fonseca, Gonçalo L. "Endogenous Growth Theory: Arrow, Romer and Lucas." History of Economic Thought Website. Accessed July 28, 2013. http://www.hetwebsite.org/het/essays/growth/endogenous.htm.

forever freedom. "My Disappointment at the Future." Longecity forum. July 26, 2007. Accessed July 28, 2013. http://www.longecity.org/forum/topic/17025-my-disappointment-at-the-future/.

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

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

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

Goertzel, Ben, and Cassio Pennachin, eds. *Artificial General Intelligence*. Cognitive Technologies. Berlin: Springer, 2007. doi:10.1007/978-3-540-68677-4.

Good, Irving John. "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, 1965. doi:10.1016/S0065-2458(08)60418-0.

Goodreads. "Epicurus Quotes." 2013. Accessed July 28, 2013. http://www.goodreads.com/author/quotes/114041.Epicurus.

Guha, R. V., and Douglas B. Lenat. "Re: CycLing Paper Reviews." *Artificial Intelligence* 61, no. 1 (1993): 149–174. doi:10.1016/0004-3702(93)90100-P.

Hall, John Storrs. "Engineering Utopia." In Wang, Goertzel, and Franklin, *Artificial General Intelligence 2008*, 460–467.

Hanson, Robin. "Britain Was Too Small." *Overcoming Bias* (blog), June 19, 2008. http://www.overcomingbias.com/2008/06/britain-was-too.html.

———. "Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Colonization." Unpublished manuscript, July 1, 1998. Accessed April 26, 2012. http://hanson.gmu.edu/filluniv.pdf.

———. "Cut Medicine In Half." *Overcoming Bias* (blog), September 10, 2007. http://www.overcomingbias.com/2007/09/cut-medicine-in.html.

———. "Dreams of Autarky." Unpublished manuscript, September 1999. Last revised September 2001. http://hanson.gmu.edu/dreamautarky.html.

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

———. "Economics of Nanotech and AI." Paper presented at Foresight 2010: the Synergy of Molecular Manufacturing and AGI, January 16–17, 2010. Powerpoint file at http://hanson.gmu.edu/ppt/Econ of AI n Nanotech.ppt. http://vimeo.com/9508131.

———. "Economics of the Singularity." *IEEE Spectrum* 45, no. 6 (2008): 45–50. doi:10.1109/MSPEC.2008.4531461.

———. "Enhancing Our Truth Orientation." In *Human Enhancement*, 1st ed., edited by Julian Savulescu and Nick Bostrom, 257–274. New York: Oxford University Press, 2009.

———. "Five Nanotech Social Scenarios." In *Nanotechnology: Societal Implications—Individual Perspectives*, edited by Mihail C. Roco and William Sims Bainbridge, 109–113. Dordrecht, The Netherlands: Springer, 2007.

———. "If Uploads Come First: The Crack of a Future Dawn." *Extropy* 6, no. 2 (1994). http://hanson.gmu.edu/uploads.html.

———. "In Innovation, Meta is Max." *Overcoming Bias* (blog), June 15, 2008. http://www.overcomingbias.com/2008/06/meta-is-max---i.html.

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

———. "Meet the New Conflict, Same as the Old Conflict." *Journal of Consciousness Studies* 19, nos. 1–2 (2012): 119–125. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00008.

———. "Morality Is Overrated." *Overcoming Bias* (blog), March 18, 2008. http://www.overcomingbias.com/2008/03/unwanted-morali.html.

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

———. "Natural Genocide." *Overcoming Bias* (blog), June 18, 2008. http://www.overcomingbias.com/2008/06/natural-genocid.html.

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

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

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

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

Haughwout, Andrew, James Orr, and David Bedoll. "The Price of Land in the New York Metropolitan Area." *Current Issues in Economics and Finance* 13, no. 3 (2008). Accessed June 21, 2013. http://www.newyorkfed.org/research/current_issues/ci14-3/ci14-3.html.

Hawks, John, Eric T. Wang, Gregory M. Cochran, Henry C. Harpending, and Robert K. Moyzis. "Recent Acceleration of Human Adaptive Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 104, no. 52 (2007): 20753–20758. doi:10.1073/pnas.0707650104.

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

Johnson, George. "Eurisko, the Computer with a Mind of Its Own." Alicia Patterson Foundation. 1984. Accessed July 28, 2013. http://aliciapatterson.org/stories/eurisko-computer-mind-its-own.

Jones, Nicola. "Middle-eastern Farmers 'Civilised' Europe." *New Scientist*, August 5, 2002. Accessed June 26, 2013. http://www.newscientist.com/article/dn2634-middleeastern-farmers-civilised-europe.html.

Kahneman, Daniel, and Dan Lovallo. "Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking." *Management Science* 39, no. 1 (1993): 17–31. doi:10.1287/mnsc.39.1.17.

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

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

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

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

Legg, Shane, and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence." *Minds and Machines* 17, no. 4 (2007): 391–444. doi:10.1007/s11023-007-9079-x.

Lettvin, Moishe. "The Windows Shutdown Crapfest." *Moishe's Blog* (blog), November 24, 2006. http://moishelettvin.blogspot.com/2006/11/windows-shutdown-crapfest.html.

Liberman, Nira, and Yacov Trope. "The Psychology of Transcending the Here and Now." *Science* 322, no. 5905 (2008): 1201–1205. doi:10.1126/science.1161958.

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

Maddison, Angus. "Measuring and Interpreting World Economic Performance 1500–2001." *Review of Income and Wealth* 51, no. 1 (2005): 1–35.

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

Markoff, John. "Computer Wins on 'Jeopardy!': Trivial, It's Not." *New York Times*, February 16, 2011. http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html.

McDaniel, Michael A. "Big-Brained People are Smarter: A Meta-Analysis of the Relationship between In Vivo Brain Volume and Intelligence." *Intelligence* 33, no. 4 (2005): 337–346. doi:10.1016/j.intell.2004.11.005.

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

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

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

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

Muehlhauser, Luke, and Chris Williamson. "Ideal Advisor Theories and Personal CEV" (2013). http : / / intelligence . org / files / IdealAdvisorTheories.pdf.

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

Norvig, Peter. "On Chomsky and the Two Cultures of Statistical Learning." May 27, 2011. Accessed July 28, 2013. http : / / norvig . com / chomsky.html.

Omohundro, Stephen M. "The Basic AI Drives." In Wang, Goertzel, and Franklin, *Artificial General Intelligence 2008*, 483–492.

Pelikan, Martin, David E. Goldberg, and Erick Cantú-Paz. "Linkage Problem, Distribution Estimation, and Bayesian Networks." *Evolutionary Computation* 8, no. 3 (2000): 311–340. doi:10 . 1162 / 106365600750078808.

Pennachin, Cassio, and Ben Goertzel. "Contemporary Approaches to Artificial General Intelligence." In Goertzel and Pennachin, *Artificial General Intelligence*, 1–30.

Ponce de León, Marcia S., Lubov Golovanova, Vladimir Doronichev, Galina Romanova, Takeru Akazawa, Osamu Kondo, Hajime Ishida, and Christoph P. E. Zollikofer. "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, no. 37 (2008): 13764–13768. doi:10.1073/pnas.0803917105.

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

Reiner, Rob, dir. *The Princess Bride*. Produced by Andrew Scheinman, written by William Goldman. 20th Century Fox, September 25, 1987. Film.

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

Rosati, Connie S. "Persons, Perspectives, and Full Information Accounts of the Good." *Ethics* 105, no. 2 (1995): 296–325. doi:10.1086/293702.

Russell, Stuart J., and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 1st ed. Upper Saddle River, NJ: Prentice-Hall, 1995.

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

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

Sandberg, Anders, and Nick Bostrom. *Whole Brain Emulation: A Roadmap*. Technical Report, 2008-3. Future of Humanity Institute, University of Oxford, 2008. http://www.fhi.ox.ac.uk/wp-content/uploads/brain-emulation-roadmap-report1.pdf.

Schopf, J. William. "Disparate Rates, Differing Fates: Tempo and Mode of Evolution Changed from the Precambrian to the Phanerozoic." *Proceedings of the National Academy of Sciences of the United States of America* 91, no. 15 (1994): 6735–6742. doi:10.1073/pnas.91.15.6735.

Shulman, Carl. "Evolutionary Selection of Preferences." Private post, *Reflective Disequilibria* (blog), November 2008. http://reflectivedisequilibria.blogspot.com/2008/11/evolutionary-selection-of-preferences.html.

———. "Zero and Non-zero-sum Games for Humans." Private post, *Reflective Disequilibria* (blog), November 2008. http://reflectivedisequilibria.blogspot.com/2008/11/zero-and-nonzero-sum-games-for-humans.html.

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

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

Sinn, Hans-Werner. "Weber's Law and the Biological Evolution of Risk Preferences: The Selective Dominance of the Logarithmic Utility Function." *Geneva Papers on Risk and Insurance Theory* 28, no. 2 (2003): 87–100. doi:10.1023/A:1026384519480.

Spinney, Laura. "The Gene Chronicles." *New Scientist*, February 7, 2004, no. 2433. Accessed June 26, 2013. http://www.newscientist.com/article/mg18124335.200.

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

Tegmark, Max. "Importance of Quantum Decoherence in Brain Processes." *Physical Review E* 61, no. 4 (2000): 4194–4206. doi:10.1103/PhysRevE.61.4194.

Tsur, Yacov, and Amos Zemel. *On Knowledge-Based Economic Growth*. Discussion Paper 8.02. Rehovot, Israel: Department of Agricultural Economics and Management, Hebrew University of Jerusalem, November 2002.

Tuomi, Ilkka. "The Lives and the Death of Moore's Law." *First Monday* 7, no. 11 (2002). http://firstmonday.org/ojs/index.php/fm/article/view/1000/921.

Vedantam, Shankar. "In Face of Tragedy, 'Whodunit' Question Often Guides Moral Reasoning." *Washington Post*, December 8, 2008. Accessed November 25, 2012. http://www.washingtonpost.com/wp-dyn/content/article/2008/12/07/AR2008120702830.html.

Wang, Pei, Ben Goertzel, and Stan Franklin, eds. *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS, 2008.

Weitzman, Martin L. "Recombinant Growth." *Quarterly Journal of Economics* 113, no. 2 (1998): 331–360. doi:10.1162/003355398555595.

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

Wiles, Andrew. "Modular Elliptic Curves and Fermat's Last Theorem." *Annals of Mathematics* 142, no. 3 (1995): 443–551. doi:10.2307/2118559.

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

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

Yudkowsky, Eliezer. "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, 2008.

———. *Coherent Extrapolated Volition*. The Singularity Institute, San Francisco, CA, May 2004. http://intelligence.org/files/CEV.pdf.

———. *Complex Value Systems are Required to Realize Valuable Futures*. The Singularity Institute, San Francisco, CA, 2011. http://intelligence.org/files/ComplexValues.pdf.

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

———. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. The Singularity Institute, San Francisco, CA, June 15, 2001. http://intelligence.org/files/CFAI.pdf.

———. "Economic Definition of Intelligence?" *Less Wrong* (blog), October 29, 2008. http://lesswrong.com/lw/vc/economic_definition_of_intelligence/.

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

———. "Excluding the Supernatural." *Less Wrong* (blog), September 12, 2008. http://lesswrong.com/lw/tv/excluding_the_supernatural/.

———. "Intelligence in Economics." *Less Wrong* (blog), October 30, 2008. http://lesswrong.com/lw/vd/intelligence_in_economics/.

———. "Levels of Organization in General Intelligence." In Goertzel and Pennachin, *Artificial General Intelligence*, 389–501.

———. "Natural Selection's Speed Limit and Complexity Bound." *Less Wrong* (blog), November 4, 2007. http://lesswrong.com/lw/ku/natural_selections_speed_limit_and_complexity/.

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

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

Yudkowsky, Eliezer. "Protein Reinforcement and DNA Consequentialism." *Less Wrong* (blog), November 13, 2007. http://lesswrong.com/lw/l2/protein_reinforcement_and_dna_consequentialism/.

———. "Reply to Holden on 'Tool AI.'" *Less Wrong* (blog), June 12, 2012. http://lesswrong.com/lw/cze/reply_to_holden_on_tool_ai/.

———. "Staring into the Singularity." Unpublished manuscript, 1996. Last revised May 27, 2001. http : / / yudkowsky . net / obsolete / singularity.html.

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

———. "The Bedrock of Fairness." *Less Wrong* (blog), July 3, 2008. http://lesswrong.com/lw/ru/the_bedrock_of_fairness/.

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

———. "Yehuda Yudkowsky, 1985–2004." November 2004. Last revised May 8, 2005. http://yudkowsky.net/other/yehuda.