ROUGH DRAFT: When Software Goes Mental: Why Artificial Minds Mean Fast Endogenous Growth

December 27, 2009

Abstract

Economic growth has so far come from human minds. The future could bring software minds: AIs designed from scratch, or human brains transferred to computer hardware. Such minds could substitute for humans in a wide range of economic activities—including the research and development that are essential to economic growth. Once minds are software products, they can be copied, accelerated, and improved by economic activity. So we will need to model growth not as given from outside, or *exogenous*, but as caused within the economy, or *endogenous*. Many endogenous growth models exist in economics as a whole, but out of the few existing models of economies with software minds, hardly any are endogenous. We review some existing literature and attempt to close the gap by adding software minds to a model of R&D based on the well-known model of Romer (1990). We focus on the speed of growth, and consider, among other things, what happens when research becomes more serial, and what happens when software minds suddenly gain access to a large fixed capital base, or "hardware overhang". Even given conservative assumptions, we find solutions that diverge, up to any ultimate limits to technology. (In a differential equations setting, if a process feeds back on itself, a "knife-edge critique" shows it generically either stagnates or blows up). If such a "hard takeoff" blowup is likely, or at least not too unlikely based on structural uncertainty, serious implications for policy follow. We conclude with some tentative comments on what these are.

1 Introduction

1.1 Modeling Long-Run Growth: Why and How

What we should do depends on what the future is like. If the future is rich, we should do one thing; if the future is poor, we should do another. If some possible futures are rich, we should try to enter those possible futures. If the future gets rich very fast, we should be prepared for upheaval. So one thing that would be good to know about the future is how fast the economy will grow. One way of finding out is by projecting recent trends: the economy has been growing at a few percent per year, corresponding to a doubling time measured in decades. Maybe we should expect it will continue to do so, with appropriate error bars.

This is the method economists have typically used, a hundred years ahead and more. For example, consider models designed to assess the damage we can expect from climate change. Nordhaus and Boyer (2003, p. 53), the authors of perhaps the most prominent such model, write: "It is probably impossible to provide accurate long-run projections given the rapid rate of social, economic, political, and institutional changes. Perhaps the best one can do is to heed the words of the eminent Harvard economic forecaster, Otto Eckstein, who advised that if we cannot forecast well, we should forecast often."

Trend extrapolation is a good way to get default mainline probability estimates. But we may turn out wildly wrong if underlying conditions change, so we're sampling from a different distribution. Taleb (2007) has called such events "black swans".

"Forecasting often" is one way to deal with such changes in conditions, but we would like our forecasts to be correct right now, so as to have time to prepare.

Economic growth was much slower before the industrial revolution than it is now, and much slower still before agriculture. Extrapolating trends from either of these eras would have yielded enormous errors, the more into the future the more so.

So if our trends could be rendered misleading by future conditions, can we find candidates for what such conditions could be?

If there were a major nuclear war, that would slow down growth down. The same is true for any disaster that crippled or destroyed human civilization, whether natural or technological.

On the other end, it's conceivable that new technologies could speed up the process of economic growth itself. Past growth already incorporates science and technology as the main drivers of economic growth, and it already incorporates several things that have made the process of scientific discovery more effective, but it may not account for genuinely new phenomena.

Our claim in this paper is that, if we developed ways to have minds on the same level as those of humans instantiated as software, that qualifies as such a trend-breaking technology.

1.2 Software Minds: a Trend-Breaking Factor?

These hypothetical future entities, which we will (somewhat artificially) refer to as "software minds", may in many ways not resemble human minds. But, by our definition, like humans, they would be able to fulfill the same roles in the economy: function as human capital in producing goods and ideas. To reiterate, the property we're interested in here is not anything like "consciousness" or "sentience": it is the ability to substitute for humans in many areas of economics, including research, which we can fairly call “general intelligence”. On the other hand, unlike humans and like programs, it would be possible to copy these minds at much less cost than the two decades needed to produce an economically or scientifically productive human.

Such entities sound like they belong in the domain of science fiction. But many experts believe they are realistic: over the next decades, we may invent technologies leading up to the development of the first software mind. We see these software minds as coming in two different possible kinds.

Artificial intelligence, as an academic field, currently focuses on individual, narrowly-defined problems. But in the future, we may achieve “strong AI” or “artificial general intelligence”: programs, designed from scratch, that have a generalized capability to solve complex problems, like the one humans were imbued with by evolution. Such programs are one kind of “software mind”.

The other kind is sometimes called a “whole brain emulation” or “brain upload” (Bostrom and Sandberg, 2008). By this we mean a program created by cutting a human brain into slices or otherwise getting at its structure, and then scanning all the structure into a computer program, which is then run, and should emulate the original brain’s behavior.

In growth economics, there are exogenous growth models and endogenous growth models. The exogenous growth models have technological progress coming from outside the model, unaffected by the variables within. The endogenous growth models have technological progress coming from some sort of research done in the model, with the amount of research done depending on the amount of economic activity and the fraction of it allocated to research by profit-maximizing agents. It is important that the growth in our model depend on variables within the model. Growth comes mainly from minds creating good ideas, so if minds themselves become a technology, continuing to model growth as if steadily given from outside would be grossly inaccurate.

1.3 Modeling with Infinities

We will sometimes talk about variables blowing up to infinity. When a model does this, it may seem like a bad feature. There are a few different reasons why economists have tended not to take such solutions seriously.

First, obviously, the past has never blown up to infinity. That might be reason to think the future will never blow up to infinity. Any model, applied to the past, must reproduce growth rates that are roughly constant over the relevant time scales or be ruled out by the evidence. (Also, Hanson (1998) has argued past economic growth is characterized by transitions between exponential growth modes; these should be accounted for, too.)

Second, there are physical limits to growth. An economy, as far as we know, can expand at most at the speed of light, which in three dimensions implies a cubic growth of volume. The number of computations that can be implemented in a given volume of space is, we suspect, bounded by the holographic entropy bound (Bousso, 2003).

Third, infinity is hard to work with. Where agents base their behavior rationally on future expectations, expectations of infinite wealth cause the model to become ill-defined even before infinity is actually achieved.

Fourth, infinity just sounds absurd. In the words of Weitzman (2009a): “There is a natural tendency to sneer at economic models that yield infinite outcomes. This reaction is presumably based on the idea that infinity is a ridiculous result; therefore any model that has an infinity symbol in it is fundamentally mis-specified, and thus dismissable.” It’s tempting to say that a model with infinite solutions anywhere is broken, and uninformative even in regimes far from where the blowups actually occur.

We still manage to feel confident about using models that generate infinities, for the following reasons. If we can identify some conditions that made it impossible for the past to blow up, and if that condition does not pertain to the future, then this renders past evidence irrelevant. We claim that the lack of software minds is such a condition. As no software intelligences existed in the past, a model with future software intelligence and a model without future software intelligence should lead to close to the same predictions for the past.

Note that models where certain parameters go to infinity are sometimes used in the physical sciences, to describe real phenomena: for example, models of the superfluid transition of helium, where the specific heat capacity goes to infinity, and viscosity goes to zero; or models of superconductivity, where conductivity goes to infinity. Even though infinity isn’t the real value, these well-established results from condensed matter physics illustrate that some variable being infinite does not imply that the model is making useless predictions. “This thing is infinite” can be shorthand for “this thing is unconstrained by the structure of our model, even if it might be constrained by something far outside its domain”.

As for your absurdity bias, this is not so much an economic as a psychological problem. Absurdity, in the wider world, is a good cue to falsehood. But if we have an idea as to the causes of absurdity and falsehood, it will be more helpful to look at truth or falsehood directly.

1.4 Overview

The rest of this paper is arranged as follows. We will first review past work that has modeled economic growth with intelligent software mathematically, in Section 2. Then, in Section 3, we try to spell out some features an economic model of the type we are building should have. In Section 3.1 we go into some detail about the novel feedback processes that should be represented in the mathematics. In Section 4, we judge past models by the criteria we found. In Section 5, we make some comments about the space of possible models, starting from a simple differential equation where the level of technology feeds back directly onto itself, then adding limiting factors. Section 6 constitutes the core of this paper. Starting from Romer's 1990 model, we try to build models of economies with software minds. We explore various assumptions. In Section 7 we explore the implications the deep uncertainties around growth curves for economies with software minds.

Since this paper depends on concepts in artificial intelligence and economics, we will not be able to assume all of our readers know the prerequisite ideas in both fields. This may mean we spent much time explaining the obvious. For this we apologize.

2 Past Models of Software Minds

A few different authors have built models for the economic implications of machine intelligence.

2.1 Kurzweil

Probably the best-known such model is Kurzweil (2001).

Kurzweil has computing power V growing as a linear function of "world knowledge" W, with the change in world knowledge linear in available computing power:

$$V = c_1 W \tag{2.1}$$

$$\dot{W} = c_2 V \tag{2.2}$$

From these two equations, Kurzweil deduces smooth exponential growth:

$$V = V_0 e^{c_1 c_2 t} \tag{2.3}$$

He extends the model by letting the amount of resources going into computing grow as a second, much slower exponential:

$$N = c_3 e^{c_4 t} \tag{2.4}$$

and then modifying the dynamics of world knowledge growth from (2.2) to get

$$\dot{W} = c_2 N V \tag{2.5}$$

which is solved by

$$V = V_0 e^{\frac{c_1 c_2 c_3}{c_4} e^{c_4 t}}. \tag{2.6}$$

So Kurzweil's model, as reflected in Kurzweil's futurism, predicts a future where growth—although eventually fast by conventional standards—is smooth, with no sudden jumps in growth rate.

2.2 Moravec

We saw that in Kurzweil's model, computing power is a linear function of world knowledge. That is, an exponential increase of computing power as per Moore's Law requires an exponential growth in world knowledge. Other specifications are possible.

Some calculations on this have been done by Moravec (1999). In his more general model, V is any function of W, though he focuses on the exponential as an example. He distinguishes two regimes: a human regime, where the change in W is proportional to time; and a machine regime, where, as in Kurzweil, the change in W is proportional to available computing power. In the full model, with both regimes pasted together, the model is:

$$V = f(W) \tag{2.7}$$

$$\dot{W} = 1 + V \tag{2.8}$$

Moravec (2003) shows that, as long as f is given by a power higher than 1 (or something even faster), solutions diverge to infinity in finite time.

That means that blowups to infinity can happen under assumptions that are not obviously false—that's it's a live option. In a sense, Kurzweil's model, in the space of all models, is at the edge of the region that doesn't blow up.

2.3 Hall

Hall (2007) offers another smooth exponential growth model. Unlike Kurzweil or Moravec, Hall explicitly distinguishes hardware and software growth.

$$Q_t = Ce^{it} \tag{2.9}$$

$$Q_t = Ce^{rpt} \tag{2.10}$$

$$Q_y = e^{(rp)(y-y_0)} = e^{0.000002(y-y_0)} \tag{2.11}$$

$$e^{0.6(y-y_0)} \tag{2.12}$$

$$Q_y = Ce^{1.2(y-y_0)} \tag{2.13}$$

Technology growth, in Hall's model, is exogenous. The number of researchers increases over time, but research output continues at its own pace independent of that number. Hence, unlike in Moravec's model, the solutions remain smooth. (Fill in more text.)

2.4 Hanson

The models discussed so far do not use standard economic growth theory: they make no attempt to model outcomes as resulting from competing rational agents. Hanson (2001) does so, by adapting the Solow-Swan model of exogenous growth. Computer hardware, M, is distinguished from ordinary capital K to account for a fast (Moore's Law) drop in hardware prices, with P the price of hardware relative to everything else. Labor L is split into human labor H, and intelligent machine labor U. The overall production function is:

$$Y = AL^{\alpha}K^{\beta}M^{\gamma} \tag{2.14}$$

Here A is a general technology factor, and returns to total input are conservatively assumed to be diminishing: $\alpha + \beta + \gamma < 1$.

Using the simplifying assumption of a constant interest rate, and taking into account various arbitrage arguments, Hanson derives a growth rate given by

$$\ln' Y = \frac{\ln' A + \alpha \ln' H - \gamma \ln' P}{1 - \gamma - \beta} \tag{2.15}$$

Hanson then demonstrates that, within his model, if we assume intelligent software allowing computer hardware to substitute for human labor, this creates a change in the economic growth rate corresponding to allowing α to go to zero and γ to go to $\alpha + \gamma$. Because computer prices fall faster than population rises, and because of the corresponding decrease in the denominator of (2.15), this results in a massive increase in growth rate. For plausible parameter values, this leads to doubling times at least a factor of ten shorter.

Investigating consequences for human wages, Hanson finds that at first, when machines complement human labor, wages go up, but when machines also start substituting for human labor, wages may go down.

A continuum of job types, some more suitable for humans and some more suitable for machines, and a hardware/software distinction, refine his model but do not significantly affect its major conclusions.

Hanson's model is exogenous, but he also briefly discusses an endogenous growth version, based on "learning by doing" (Solow, 1997). For certain parameter values, his model again outputs fast exponential growth. For other parameter values, his model has no steady exponential growth solutions at all.

So Hanson's model hints at the possibility, which we will argue for further in this paper, that endogenous growth models of software minds support explosive behavior.

2.5 Johansen & Sornette

Johansen and Sornette (2001, section 5.3) present, as part of a paper arguing for a future finite-time singularity in population and economic indicators, a class of endogenous technology growth models of a different kind. (This model is presented as part of a paper extrapolating a future finite-time singularity in population and economic indicators based on "log-periodic oscillations". The paper has been criticized as an example of pathology in econophysics (Gallegati et al., 2006), but this is not relevant to the current discussion.) They have economic output depending on a combination of capital, labor, and existing technology:

$$Y(t) = [(1-a_K)K(t)]^{\alpha}[A(t)(1-a_L)L(t)]^{1-\alpha} \tag{2.16}$$

Further, they have the rate of technology change, too, depending on a combination of capital, labor, and existing technology:

$$\frac{dA}{dt} = B[a_K K(t)]^{\beta}[a_L L(t)]^{\gamma}[A(t)]^{\theta}, \qquad B > 0,\ \beta \geq 0,\ \gamma \geq 0 \tag{2.17}$$

For the growth of capital, this implies:

$$\frac{dK}{dt} = sY(t) = s[(1-a_K)K(t)]^{\alpha}[A(t)(1-a_L)L(t)]^{1-\alpha} \tag{2.18}$$

The section explores a few different specifications for the rate of growth of labor and capital, amounting to the addition of different feedback loops. For example, if both capital and labor are held constant, the only thing left to cause potential divergences is a strong feedback in technology itself, according to an equation of the form:

$$\frac{dp}{dt} = r[p(t)]^{1+\delta} \tag{2.19}$$

If labor is held constant but capital grows as determined by a constant saving rate, the model reduces to:

$$\frac{dA}{dt} = bA^{\theta}K^{\beta} \tag{2.20}$$

$$\frac{dK}{dt} = aA^{1-\alpha}K^{\alpha} \tag{2.21}$$

and there is potential for divergences for a greater range of possible parameter values. Feedback loops can cause the system as a whole to blow up even if the individual components do not have strong feedbacks in themselves.

[latex equations in case the preceding needs to be illustrated:)

$$A(t) = A_0(t_c - t)^{-\delta} \tag{2.22}$$

$$K(t) = K_0(t_c - t)^{-\kappa} \tag{2.23}$$

$$\delta = \frac{1+\beta-\alpha}{(1-\alpha)(\theta+\beta-1)} \tag{2.24}$$

$$\kappa = \frac{2-\theta-\alpha}{(1-\alpha)(\theta+\beta-1)} \tag{2.25}$$

Finally, if labor always grows to a fixed fraction of capital (as Kremer (1993) assumes in the context of past population growth $\frac{Y(t)}{L(t)} = \bar{y}$, where the interpretation is people growing up to the limits of the infrastructure supporting them), the model turns into:

$$\frac{dA}{dt} = a'[L(t)]^{\beta+\gamma}[A(t)]^{\theta}, \qquad a' > 0,\ \beta \geq 0\ \gamma \geq 0 \tag{2.26}$$

$$\frac{dL}{dt} = b'L(t)[A(t)]^{1-\alpha} \tag{2.27}$$

$$\delta = \frac{1}{1-\alpha} \tag{2.28}$$

$$\kappa = \frac{2-\theta-\alpha}{\beta+\gamma} \tag{2.29}$$

Here, the interplay between population, capital, and technology growth causes blowups for a much wider range of parameters still. Johansen & Sornette do not give an interpretation of their model in terms of artificial intelligence, but seeing this specification for the growth of labor as coming from investment into human-equivalent artificial intelligence seems natural.

2.6 Jones

Jones (2009) discusses economic growth with software intelligence based on a semi-endogenous growth model by Jones (1995). Such models are described as having "semi-endogenous growth" because returns to investment in technology diminish the more technology already exists, and so permanent increases in such investment only cause the diminishing returns to be hit earlier. In Jones's model, which among other things can be seen as a special case of the model in Johansen and Sornette (2001), technology A is described by:

$$dA/dt = wL^{a}A^{b} \tag{2.30}$$

In this context, the problem that remains is to state how software minds would change a and b.

So it would seem that there are many different ways to model growth given software minds, that give different conclusions. Can we exclude any of them as being unrealistic? Before we start building our own models, we will first offer some features we believe such a model should have.

3 Desiderata for Models of Software Minds

Current economic models depend on various assumptions. These are realistic for the economic past, present, and near future. However, some would break down under certain future technologies. Software intelligence, especially, could remove a few fixed bottlenecks and, by doing so, create new feedback loops.

This makes proper economic modeling of futures containing software minds difficult. A model, to be accurate and helpful, should ideally satisfy a number of desiderata—although we do not claim our model will satisfy all of these.

1. The model should give sensible results in the special case where the number of software minds is always zero. It must be consistent with our past experience of roughly exponential growth.
2. The assumptions about parameters and functional forms should not themselves be wildly unrealistic.
3. The model should be robust: its conclusions should not depend sensitively on small parameter changes. As it is hard to get exact parameter changes from data, we should not expect the conclusions of models that are not robust to carry over to the real world.
4. Where possible, the model should have resource allocation arising from the decisions of agents. This offers some more realism, and allows us to identify extreme cases, where [...].
5. The model should not neglect any limiting factor we can expect to constrain progress.
6. The model should incorporate all the important feedback effects software minds bring.

Desiderata 4 and especially 6 suggest we should prefer endogenous growth models to exogenous growth models.

3.1 Feedback Effects from Software Minds

Desideratum 6 could use expansion: what are these important feedback effects software minds are supposed to create?

Without software intelligence, we're used to modeling the situation more or less as follows.

Some part of the economy creates more physical capital. The rate at which it does so depends on a few things. One is the level of technology. Another is the level of different inputs: human labor, human capital, physical capital. These in turn depend on how much of them exists, and on how much of them gets invested by consumers weighing present against future consumption.

Human labor is limited: populations grow, but not by orders of magnitude over the time scales we are interested in. People acquire more skills over their lifetime, but they also die. So we can see human capital as fixed, and to the extent that the same amount of education becomes more effective, count this under “technology”.

The stock of capital is not fixed, and the accumulation of capital is a cause of growth. But unless the number of people to operate the machines can keep up, diminishing returns mean growth (at least purely from capital accumulation) stagnates.

Some other part of the economy creates more knowledge. The rate at which it does so depends on a few other things. The current level of knowledge affects how easy it is to create more. The inputs matter again: labor, physical capital (“lab equipment”), human capital in the form of researchers.

Again, human labor and human capital are limited and physical capital is subject to strongly diminishing returns. Since technology has not led to explosive growth in the past [...].

With software minds, or certain other future technologies, several effects can arise.

1. The stock of unskilled labor—roughly, human bodies and robots—stops being limited, and starts being added to by the production sector. The field of robotics, while distinct from human-level artificial intelligence, may advance along with it. Artificial wombs or the removal of limits to the growth of human population may have the same effect, but more slowly. If labor can keep accumulating along with capital, returns to reinvested output diminish less quickly.

2. The stock of skilled labor—roughly, human brains and software minds—stops being limited and starts being added to by the production sector. Intelligent programs could substitute for humans in jobs requiring intelligence; the ease of copying and selecting make training cheaper and skill greater, the more so with better software mind technology. Again, other technologies could have the effect of creating more humans. Again, returns to reinvested output would diminish less quickly.

   If both 1 and 2 are in play, the loop is fully closed: all the inputs to production are now also outputs. If a fixed fraction of output is reinvested, then even with no technological progress, exponential growth results. (Consider self-replicating Von Neumann machines.)

3. The number of scientists stops being fixed. If new researchers are software minds, producing them requires only that copies of the same design be run on more hardware.

4. The speed at which scientists think stops being fixed. Even keeping the number of computations constant, a greater serial speed should increase research efficiency. This is because serial computations can be sacrificed for parallel computations, but not the other way around: problems requiring long chains of logical dependencies require serial speed. So here, research done feeds into the efficiency of future research.

5. The intelligence of scientists stops being fixed. Even keeping the number and serial speed of computations constant, we can expect qualitatively better algorithms to increase the efficiency of software minds doing research. If computing power were all that mattered, whales would be able to outrace human researchers. If they have, they have thus far been secretive about their findings.

Multiple such effects are likely to come into play at the same time, and build on one another. As we saw in Johansen and Sornette (2001), feedback effects that are individually too weak can combine to cause explosive growth. Of course, a given unit of extra output can't simultaneously feed back to create a unit of unskilled labor, a unit of skilled labor, and a researcher; if a fixed fraction of everything is reinvested into everything, allowing more kinds of things to reinvest in decreases what this fraction is. But to the extent that increases in different inputs complement each other, returns still diminish less quickly. (In reality, the fractions aren't fixed, but depend on how much profit agents can expect to capture from investing in one kind of inputs versus another.)

4 Critique of Past Work

Kurzweil's model has several problems.

It is not clear why the growth of world knowledge should be proportional to computing power. In a world where humans do most of the research, giving a scientist twice as many computers does not double that scientist's output. (A literal interpretation of the model would entail that the world's research output doubled when people built the world's second computer.)

In a world where software minds do most of the research, the assumption makes more sense. Even so, the model of science implied here is too simple. It does not take into account changes to the software that minds run on. [...]

Kurzweil further assumes computing power, V, is a linear function of world knowledge, W. This is an extremely conservative assumption: it implies that a fixed base of hardware researchers would need twice as long to invent each subsequent doubling of computing power, instead of a constant time as in Moore's law.

As we saw in Moravec's model above, any power higher than 1 in this function would cause a blowup in finite time. If we see the space of all models as divided into a region where solutions blow up, and a region where it does not, then Kurzweil's model is exactly at the boundary. That means it fails the requirement of robustness.

Moravec's model, while in one way more general than Kurzweil's, is still simple. [...]

Hall's model has a different problem. His model features, as a result of exogenous hardware and software improvements, an exponentially growing number of software mind researchers. But these researchers [...]

Hanson's model is, as far as we know, the only model for the development of software minds to use standard economic growth models. There are a few reasons why, nonetheless, we believe its growth projections are underestimates.

First, there is the conservative assumption of diminishing returns to total input; that is, $\alpha + \beta + \gamma < 1$. Typically, returns are assumed constant, or $\alpha + \beta + \gamma = 1$. This can be justified by a "replication argument": two copies of the same economy that didn't interact would together create twice as much output.

Second, the main part of the model assumes growth is exogenous. Intelligent software minds are created at an exponential rate, but are only ever applied in production, never becoming researchers. As argued before, this ignores some of the important feedback loops that would arise from software mind technology.

Section 4 of the paper does discuss an endogenous growth model, based on an early endogenous growth model by Arrow called "learning by doing" (Solow, 1997). This model assumes that the rate of change of technology is proportional to some power of total economic output. Moreover, it assumes this power is smaller than 1, making this actually a "semi-endogenous growth model". (The relative price of computing power, P, behaves similarly.) For the model to output exponential growth rather than divergence in finite time, some rather restrictive conditions on these parameters have to hold—low-hanging fruit must get depleted very quickly, returns to total output must be very diminishing, or preferably both.

Jones's model

Johansen & Sornette's model

5 Linearity and Limits

Simple differential equation for feedback loops:

$$\frac{dA}{dt} = A^{\alpha} \tag{5.1}$$

Blows up or stagnates or is exactly exponential, with measure zero. This has been called the knife-edge critique in the literature, see for example Solow (2000). Briefly discuss Growiec (2007), Growiec (2008), Dalgaard and Kreiner (2003). Put graphs from mathematica here, linear and log scale. If it's bounded above zero and below infinity then it does end up looking exponential, so it's not really measure zero in the wider space. Cite the paper that makes that point. Maybe also put in some graphs about a model where α isn't constant. Discuss relation to Yudkowsky's "strong self-improvement".

Differential equation for feedback loops with limits:

$$\frac{dA}{dt} = A^{\alpha}(L - A)^{-\beta} \tag{5.2}$$

Logistic equation as special case. If β high, sudden bump into ceiling analogous to hard takeoff (seems plausible). Probably go into more detail here about physical limits to growth. Put graphs from mathematica here, linear and log scale. Also alternative equation with higher-hanging fruit taking exponentially more resources, graphs.

Semi-endogenous growth is a sort of limit. That's where tech growth is less than linear in resources expended because of low-hanging fruit getting depleted. Called semi-endogenous growth because putting more resources into R&D only delays the point where you hit high-hangingness of fruit. Briefly discuss Young (1998), Dalgaard and Kreiner (2001), Jones (1995), Segerstrom (1998), Howitt (1999). Note that semi-endogenous growth models can still blow up with the right kind of other factors, like population growth, as in Johansen and Sornette (2001).

According to Ha and Howitt (2007) endogenous growth and semi-endogenous growth don't account as well for the evidence as Schumpeterian growth, which says increasing R&D inputs are counteracted by expanding product varieties. Link to empirical estimates. No obvious continued growth in such product varieties given AI, especially AI that likes fast long-run growth more than instant profits. Limits to how small a fraction you can ultimately spend on basic research.

6 Romer With Robots

Big question: do we work off the original Romer (1990) model, or off a simplified version like Johansen and Sornette (2001)? Leaning toward the latter: modeling the details of consumer preferences is mathematically very hairy and relatively unenlightening; however, note Steve Rayhawk's optimal control work about behavior toward infinity under assumptions amounting to maximizing fraction of future lightcone. In a simplified model we could try out more changes and make conceptual points. Working based on the full model would probably be more publishable.

Probably paste in quick intro to Romer's model, partially excludable nonrival goods, and so on.

Add factors for serial speed, qualitative smartness. If there are diminishing returns to parallel research (Brooks (1995)) at each time step, it can be shown this leads to a higher exponent the more you replace parallel by serial computing power. Paste in math here. This is only partly true if only part of the research is serializable. Paste in math about serializable theory with parallel experiments.

Discuss models with time delays. Paste in quick proof they can't blow up to infinity. Fixed time delay unrealistic: the more it starts bothering you the more you can trade the shortness of these time steps off for other things. Endogenous time delays.

Discrete time models where there are N human time steps in each AI time step.

Maybe model hardware overhangs somehow, that is to say sudden availability of source of hardware that agents being modeled didn't expect.

Maybe model repurposability of hardware.

Maybe model extreme ”blowup in brain in box in basement” scenario.

What about James D. Miller's argument that widespread awareness of the singularity will decrease investment in favor of consumption? Needs model of consumers. Agents that don't decrease their investment win. Kelly criterion, logarithmic utility functions.

Maybe model other things. Mostly it depends on what the model we end up using (full Romer or simplified Romer) ends up being congenial to.

7 Implications of Structural Uncertainty

Finally, we return to policy implications. Why does the speed of AI takeoff matter? Is this an issue that can wait until the far future, or does it affect our decisions here and now?

We should first discuss a methodological point. As argued above, there are many models of growth given AI, making a wide range of predictions, some of which are extreme. This situation mirrors that in more prominent issues such as climate change and financial risk. What is the right way to analyze costs and benefits here?

Many respond to model uncertainty by choosing a most plausible or most central model to guide their decisions. In the case of AI growth, this amounts to focusing only on plans suggested by one scenario: business as usual, stagnation, a smooth “soft takeoff”, or a sudden “hard takeoff”.

However, this method is not optimal. Decision theory (Von Neumann and Morgenstern, 1944) describes the behavior of a rational agent in terms of maximizing expected utility. Such an agent will try to achieve the best results across possible future outcomes, taking into account both the probability of different scenarios and the severity of the implied costs and benefits. “Dutch book” results show those who deviate from decision theory end up making self-defeating decisions in some situations.

The less certain one's favored model, the more important it becomes to consider out-of-model events. If the cost of a policy is small, and the benefit given some such event is large enough to overcome the probability penalty, decision theory recommends the policy. This is relevant to topics like insurance, flood engineering, and the choice of safety margins in civil engineering.

Representing parameter uncertainty solves part of the problem, but not all. Our uncertainty concerns not just parameter values, but what parameters to include, how to relate them to each other, and what functional forms to use. Some authors have called this “structural uncertainty”.

Weitzman (2009b) makes a point about structural uncertainty in the context of the economics of climate change. The argument is that if certain mathematical assumptions hold, structural uncertainty causes the expected cost to be dominated by a fat tail of improbable but extreme models that limited evidence cannot fully exclude. (In the extreme case, this could lead to the recommendation to focus only on minimizing existential risks, which Bostrom and Bostrom (2002) defines as events that would "annihilate Earth-originating intelligent life or permanently and drastically curtail its potential").

Ord et al. (2008) discuss the impact of errors in analyses on the probability of unlikely but extreme risks. Sometimes, when the best analysis places a near-zero probability on some risk, most of the risk's probability comes from cases where the best analysis is wrong. For example, if we calculate a 1 in 10^{10} chance of an asteroid impact, but there is a 1 in 100 chance that our calculation contains a mistake, it is not obvious that we should be reassured.

How does structural uncertainty play out in our own case, that of uncertain takeoff speed of technological and economic growth after the arrival of AI or brain emulations?

Suppose this paper's thesis holds up and hard takeoff scenarios have significant weight. Then, even if other models are more central or better-supported, caution justifies taking measures that have reasonable costs and chances to positively affect a hard takeoff outcome. It seems likely that such measures would exist: the potential payoff is roughly as large as the entire economy, and (to forestall "Pascal's Wager" objections) the probability of hard takeoff is, we have argued, not tiny.

What, then, could such measures look like, given that the problem will probably not become imminent until decades from now?

We can rule out some classes of plans as insufficient to mitigate hard takeoff risk. This includes all plans that require timely response to information from intermediate steps on the AI growth curve. ("If they start rebelling, we can just pull the plug.") We cannot confidently assume a hard takeoff will leave enough reaction time.

On the most general level, our conclusions suggest the returns from research into hard takeoff scenarios are high. If we could find policies that gain us safety across possible hard takeoff outcomes, these would be extremely valuable.

The economic models we used assume property rights always remain stable. If the power differentials created are as great as analyses such as this paper suggest, this assumption is unlikely to be realistic. Even slight differences in goals between machine intelligences and their owners give the former an incentive to use the wealth available to them toward their own ends. Gaps in technology, intelligence, numbers, and wealth will at some point give them the ability.

If there is no time to influence hard takeoff trajectories much from outside the software minds or whatever controls them, the outcome will depend on the goals of the minds taking off. In both the cases of artificial intelligence and brain emulations, the motivations of the minds to be copied and enhanced, and whether these motivations are preserved under the dynamics of the takeoff, determine how the new technology and wealth will be used.

Some have claimed it is both possible and desirable to build artificial intelligences whose goals are well-understood (on some level of abstraction) and predictably stable under changes to their programming. If so, the problem of hard takeoff risks may have a reliable solution along those lines.

It would also be helpful to get a better understanding of when hard takeoff disasters might happen. More information on general timing would be of use, but we should expect to have wide confidence bounds given the difficulty of predicting the future and the biases involved in thinking about these issues (Yudkowsky, 2008). If we learn more about the circumstances under which AI and brain emulation scenarios lead to hard takeoff disasters, we can better avoid setting in motion sequences of events that could end in such disasters.

Kahneman and Lovallo (1993) have distinguished between "inside view" reasoning, based on the details of a specific case, and "outside view" reasoning, based on statistics about past cases with similarities. In contexts such as project completion time, outside views often make more reliable predictions than inside views. But there are problems with taking an outside view of our case. It is not clear what the relevant class of past situations is. One can take past cases of fast speedup of economic growth, such as agriculture and the industrial revolution, but given the possibility that the AI problem is fundamentally different in structure, the outside view seems limited even if useful. And if the certainty available from outside views is limited, the arguments we have given here apply.

8 Conclusion

Don't forget to quickly discuss some other papers here, Ray Kurzweil and Vinge (1999), Powell et al. (2009), Croix and Licandro (1999), Weitzman (1998), Jones (2005), Tsur and Zemel (2002), Groth and Schou (2002).

References

Nick Bostrom and Dr. Nick Bostrom. Existential risks - analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 9, 2002.

Nick Bostrom and Anders Sandberg. Whole brain emulation: a roadmap. Technical Report 2008-3, Future of Humanity Institute, Oxford University, 2008.

Raphael Bousso. Light sheets and Bekenstein's entropy bound. *Phys. Rev. Lett.*, 90(12): 121302, Mar 2003. doi: 10.1103/PhysRevLett.90.121302.

Frederick P. Brooks, Jr. *The mythical man-month (anniversary ed.).* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995. ISBN 0-201-83595-9.

David de la Croix and Omar Licandro. Life expectancy and endogenous growth. *Economics Letters*, 65(2):255–263, 1999. URL http://econpapers.repec.org/RePEc:eee:ecolet:v:65:y:1999:i:2:p:255-263.

Carl-Johan Dalgaard and Claus Thustrup Kreiner. Is declining productivity inevitable? *Journal of Economic Growth*, 6(3):187–203, September 2001. URL http://ideas.repec.org/a/kap/jecgro/v6y2001i3p187-203.html.

Carl-Johan Dalgaard and Claus Thustrup Kreiner. Endogenous growth: A knife edge or the razor's edge? *Scandinavian Journal of Economics*, 105(1):73–86, 03 2003. URL http://ideas.repec.org/a/bla/scandj/v105y2003i1p73-86.html.

M. Gallegati, S. Keen, T. Lux, and P. Ormerod. Worrying trends in econophysics. *Physica A Statistical Mechanics and its Applications*, 370:1–6, October 2006. doi: 10.1016/j.physa.2006.04.029.

Christian Groth and Poul Schou. Can non-renewable resources alleviate the knife-edge character of endogenous growth? *Oxford Economic Papers*, 54(3):386–411, July 2002. URL http://ideas.repec.org/a/oup/oxecpp/v54y2002i3p386-411.html.

Jakub Growiec. Beyond the linearity critique: The knife-edge assumption of steady-state growth. *Economic Theory*, 31(3):489–499, June 2007. URL http://ideas.repec.org/a/spr/joecth/v31y2007i3p489-499.html.

Jakub Growiec. Knife-edge conditions in the modeling of long-run growth regularities. MPRA Paper 9956, University Library of Munich, Germany, July 2008. URL http://ideas.repec.org/p/pra/mprapa/9956.html.

Joonkyung Ha and Peter Howitt. Accounting for trends in productivity and r&d: A schumpeterian critique of semi-endogenous growth theory. *Journal of Money, Credit and Banking*, 39(4):733–774, 06 2007. URL http://ideas.repec.org/a/mcb/jmoncb/v39y2007i4p733-774.html.

J. Storrs Hall. *Beyond AI: Creating the Conscience of the Machine*. Prometheus Books, 2007. ISBN 1591025117.

Robin Hanson. Economic growth given machine intelligence. *Journal of Artificial Intelligence Research*, 2001.

Robin Hanson. Long-term growth as a sequence of exponential modes. In *George Mason University*, pages 9–3, 1998.

Peter Howitt. Steady endogenous growth with population and r &amp; d inputs growing. *Journal of Political Economy*, 107(4):715–730, August 1999. URL http://ideas.repec.org/a/ucp/jpolec/v107y1999i4p715-730.html.

Anders Johansen and Didier Sornette. Finite-time singularity in the dynamics of the world population, economic and financial indices. *Physica A*, 294(3-4):465–502, May 2001.

Charles I. Jones. Growth and ideas. In Philippe Aghion and Steven Durlauf, editors, *Handbook of Economic Growth*, volume 1, Part B, chapter 16, pages 1063–1111. Elsevier, 1 edition, 2005. URL `http://econpapers.repec.org/RePEc:eee:grochp:1-16`.

Charles I Jones. R&d-based models of economic growth. *Journal of Political Economy*, 103(4):759–84, August 1995. URL `http://ideas.repec.org/a/ucp/jpolec/v103y1995i4p759-84.html`.

Garett Jones. Artificial intelligence and economic growth: a few finger-exercises, 2009. URL `http://mason.gmu.edu/~gjonesb/AIandGrowth`.

Daniel Kahneman and Dan Lovallo. Timid choices and bold forecasts: a cognitive perspective on risk taking. *Manage. Sci.*, 39(1):17–31, 1993. ISSN 0025-1909. doi: http://dx.doi.org/10.1287/mnsc.39.1.17.

Michael Kremer. Population growth and technological change: One million B.C. to 1990. *The Quarterly Journal of Economics*, 108(3):681–716, August 1993. URL `http://ideas.repec.org/a/tpr/qjecon/v108y1993i3p681-716.html`.

Ray Kurzweil. The law of accelerating returns. 2001. URL `http://www.kurzweilai.net/articles/art0134.html?printable=1`.

Hans Moravec. Simple equations for Vinge's technological singularity. 1999. URL `http://www.frc.ri.cmu.edu/~hpm/project.archive/robot.papers/1999/singularity.html`.

Hans Moravec. SIMPLER equations for vinge's technological singularity. 2003. URL `http://www.frc.ri.cmu.edu/~hpm/project.archive/robot.papers/2003/singularity2.html`.

William D. Nordhaus and Joseph Boyer. *Warming the World: Economic Models of Global Warming*. The MIT Press, 2003. ISBN 0262640546.

Toby Ord, Rafaela Hillerbrand, and Anders Sandberg. Probing the improbable: Methodological challenges for risks with low probabilities and high stakes, 2008. URL `http://www.citebase.org/abstract?id=oai:arXiv.org:0810.5515`.

A. Powell, S. Shennan, and M. G. Thomas. Late Pleistocene Demography and the Appearance of Modern Human Behavior. *Science*, 324:1298–, June 2009. doi: 10.1126/science.1170165.

Hans Moravec Ray Kurzweil and Vernor Vinge. Singularity math trialogue. 1999. URL `http://www.kurzweilai.net/articles/art0151.html?printable=1`.

Paul M. Romer. Endogenous technological change. *The Journal of Political Economy*, 98(5): S71–S102, 1990. URL `http://dx.doi.org/10.2307/2937632`.

Paul S Segerstrom. Endogenous growth without scale effects. *American Economic Review*, 88(5):1290–1310, December 1998. URL http://ideas.repec.org/a/aea/aecrev/v88y1998i5p1290-1310.html.

Robert M. Solow. *Growth Theory: An Exposition.* Oxford University Press, USA, 2000. ISBN 0195109031. URL http://www.amazon.com/Growth-Theory-Exposition-Robert-Solow/dp/0195109031%3FSubscriptionId%3D0JYN1NVW651KCA56C102%26tag%3Dtechkie-20%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0195109031.

Robert M. Solow. *Learning from "learning by doing" : lessons for economic growth / Robert M. Solow.* Stanford University Press, Stanford, Calif. :, 1997. ISBN 0804728402 0804728410 0804728410. URL http://www.loc.gov/catdir/toc/cam027/96031847.html.

Nassim Nicholas Taleb. *The Black Swan: The Impact of the Highly Improbable.* Random House, 2007. ISBN 1400063515. URL http://www.amazon.com/Black-Swan-Impact-Highly-Improbable/dp/1400063515%3FSubscriptionId%3D0JYN1NVW651KCA56C102%26tag%3Dtechkie-20%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D1400063515.

Yacov Tsur and Amos Zemel. On knowledge-based economic growth. Discussion Papers 14997, Hebrew University of Jerusalem, Department of Agricultural Economics and Management, 2002. URL http://ideas.repec.org/p/ags/huaedp/14997.html.

John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior / by John Von Neumann and Oskar Morgenstern.* Princeton University Press, Princeton :, 1944.

Martin Weitzman. Reactions to the Nordhaus critique, 2009a. URL http://www.economics.harvard.edu/faculty/weitzman/files/ReactionsCritique.pdf.

Martin L Weitzman. On modeling and interpreting the economics of catastrophic climate change. *The Review of Economics and Statistics*, 91(1):1–19, 06 2009b. URL http://ideas.repec.org/a/tpr/restat/v91y2009i1p1-19.html.

Martin L. Weitzman. Recombinant growth. *The Quarterly Journal of Economics*, 113(2):331–360, May 1998. URL http://ideas.repec.org/a/tpr/qjecon/v113y1998i2p331-360.html.

Alwyn Young. Growth without scale effects. *Journal of Political Economy*, 106(1):41–63, February 1998. URL http://ideas.repec.org/a/ucp/jpolec/v106y1998i1p41-63.html.

E. Yudkowsky. *Cognitive biases potentially affecting judgement of global risks*, pages 86–+. 2008.