A Tale Of Two Transitions

By Robin Hanson

In this chapter I'm going to consider and compare two quite different scenarios for a future transition to a world dominated by machine intelligence. While these aren't the only possible scenarios, they are the two scenarios that I consider most likely.

One scenario is based on continuing on our current path of accumulating many small pieces of better software. The other scenario is based on the future arrival of an ability to fully emulate human brains. The first scenario is gradual and anticipated. In it, society has plenty of warnings about forthcoming changes, and time to adapt to those changes. The second scenario, in contrast, is more sudden, and potentially disruptive.

A History Of Tools

But before we get to those scenarios, let us set some context by first reviewing some basic facts about tools and machines. Between the farming revolution of around ten thousand years ago until the industrial revolution of around 1700, the world economy grew rather steadily, doubling about every millennium. While some of that growth was due to the accumulation of better tools, most of that growth was due to the accumulation of better-domesticated plants and animals. Since the industrial revolution, however, the economy has doubled about every fifteen years, and most of that growth has been due to the accumulation of better physical tools and ways to use them.

The world economy contains a great number of tasks that need to be done, and an even larger number of tasks that it would be nice to do, if only they could be done cheaply enough. When our tools were bad, almost all useful tasks had to be done by people or animals, if they were to be done at all. But as our tools have improved, those tools have displaced people (and animals) on many tasks.

Over the last century or so, our rate of displacing of people by tools on tasks has been relatively slow and steady. Also steady has been our rate of making tools better do the tasks they do better, and of doing new useful tasks that we could not previously afford to do. Some of these new tasks have been done by tools, others by people.

The reason that gains have mostly been slow and steady is that each innovation usually only improves each tool by a small amount, and improvements in different tools are largely independent of one another (Sahal 1981; Acemoglu 2009). While we have sometimes developed "general purpose tools" like computers, most gains in those have been incremental, as has progress in applying them to particular problems. Our overall economic rates of task displacement, improvement, and addition did not change greatly after the introduction of computers, even though basic abilities have been improving more rapidly for computers than for other machines.

These trends make sense if there are many kinds of physical and mental abilities, if each task requires many supporting abilities, and if the ability levels required to achieve particular tasks vary over huge ranges. Some tasks that people do are very easy, requiring low levels of many abilities, while other tasks are very hard, requiring high levels of many abilities. As the abilities of our tools improve faster than do human abilities, tools slowly rise in their relative ability to do tasks, and periodically some tools rise above the threshold of being cost-effective relative to humans, or relative to not doing the task at all.

Simple economics says that total world income is paid out to all the tasks doers, in rough proportion to the importance of each task to the outcomes we care most about, and also in proportion to the difficulty of improving the performance of each task by devoting more resources to it. The more tasks that are done, and done better, the more world income there is to pay out. Thus we can track the total value that tools and ways to use them give by the rise in world income, and we can track the relative importance of tools and people in the this economy by tracking the fraction of income given to those who own tools, relative to the fraction given to people who sell their time to earn wages.

When we track changes this way, what we see is that while tools have helped us to greatly increase the total value that we achieve by doing all of the tasks that we do, the fraction of that value that is paid to tools at any one time is still small compared to the fraction paid to people. The main reason for this is that it tends to be much faster and easier to increase the number of tools of any one time, relative to the number of skilled people. So the value that the last tool contributes to a task is still much less than the value that the last person contributes.

Today human workers directly get about 52% of all income, a figure that is down from 56% forty years ago (Karabarbounis & Neiman 2013). But the rest of income mostly does not go to tools. For example, only a few percent goes to computer-based tools. A lot of income goes to those who own valuable things like real estate, patents, and firms, things that are neither tools nor people working, but which were created in the past by people using tools.

When Will Tools Top Humans?

One plausible future scenario is that we will continue to improve our tools in the same sort of ways that we've improved them over the last century or so. That is, there will be no big sudden breakthrough that will allow big fast gains in tool effectiveness across a wide range of capabilities and tasks. There is no simple grand "theory of tools" or "theory of intelligence" that once discovered allows our tools to quickly become much more productive. In this scenario, "intelligence" is just the name we give to our ability to do well at great many mostly unrelated mental tasks.

Given this scenario, eventually there should come a point in time where tools exceed humans on almost all relevant abilities, making tools cheaper than humans for almost all tasks. At that point, the speed limits we face on making more skilled people would no longer limit the growth rate of the economy. This is because only tools would be needed to grow the economy further, and factories could quickly crank our more such tools. A tool-dominated economy might then double every month, or even faster.

How close are we today to such a transition point? One clue comes from the field of artificial intelligence [AI] research. This is the field where researchers directly and explicitly try to write software that can do hard mental tasks, tasks where tools have not yet displaced humans.

AI researchers have had substantial success over many decades. And some observers, including some AI experts, have even said that they expect that AI software with broad human level abilities will appear within a few decades (Armstrong & Sotala 2012). Often such optimism is based on hopes that we will soon discover a great powerful "architecture" or theory of intelligence, or that recent exciting developments show the early stages of such a revolution (Hanson &Yudkowsky 2013).

However, AI experts tend to be much less optimistic when asked about the topics where their estimates should be the most reliable: the rate recent of progress in the AI subfields where they have the most personal expertize. I was a professional AI researcher for nine years (1984-1993), and when I meet other such experienced AI experts informally, I am in the habit of asking them how much progress they have seen in their specific AI subfield in the last twenty years. They typically say that in that time they have only seen 5-10% of the progress required to achieve human level abilities in their subfield. They have also typically seen no noticeable acceleration over this period (Hanson 2012).

I think it is reasonable to expect that, if there is no grand theory of intelligence to discover, past rates of progress will be predictive of future rates of progress. And at those past rates, it should take two to four centuries for the typical AI subfield to reach human level abilities. Furthermore, since there would be variation across AI subfields, and since displacing humans on almost all tasks probably requires human level abilities in most AI subfields, a broadly capable human level AI would probably take even longer than two to four centuries to achieve.

In addition, in many areas of computer science software gains have closely tracked hardware gains over the last half-century, suggesting that hardware gains are important enablers of enable software gains (Grace 2013). This is a bad sign because traditional "Moore's law" trends in hardware gains have slowed lately, and are expected to slow even more over the coming decades (Esmaeilzadeh, et. al. 2012). These facts together suggest it will take even longer for our software tools to become cheaper than humans on almost all mental tasks.

This scenario, where tools continue for centuries more to improve at rates comparable to or less than those in the last century, until finally tools are cheaper than humans on almost all tasks, is the slow gradual first scenario I mentioned at the start of this chapter. In this scenario the gains in our abilities on different kinds of tasks do not depend much on each other. There is no grand theory of tool-making or intelligence that, once discovered, allows tools to improve greatly simultaneously across a wide range of tasks. Instead, different tools would mostly develop independently, getting slowly and steadily better at their various tasks, as they have for centuries.

A Slow Gradual Transition

We can use basic social science to foresee many aspects of how a transition to a tool dominated world would play out how in this scenario.

First, the world would get lots of warning about the upcoming transition to a world dominated by machine intelligence. On the one hand, growth rates may not start to increase much until the human wage fraction of income falls below five percent or even lower. On the other hand, the human wage fraction of world income may start to gradually fall a half century to a century or even longer before such a transition point. There should thus be many decades of familiar growth rates with direct vivid lessons to help humans to get used to the idea that they will eventually need to own something other than their ability to work in order to have an income.

As the fraction of world income that went to working humans fell slowly from twenty to ten to five to zero percent, individuals would decide how to diversify their personal portfolios in order to assure themselves a future income, and societies would decide how much to help the initially poor as well as those who invested poorly. Societies would also adapt their finance, law, governance, and other institutions to the new distributions of income and power.

This would be a world where total production results from a very large set of tools that are mostly improving independently, and where design changes typically only result in small changes to local functionality. These aspects of this world would greatly limit the size and scope of damages resulting from poorly considered design changes. Such mistakes would almost always be reversible at a modest cost; very few changes would have a non-zero chance of destroying society as a whole.

As is true today, some key design choices in control and governance would have larger scopes, and thus more potential for large broad damage. And this would continue to be true during and after a transition to a world dominated by machine intelligence. But it isn’t clear that the transition period would have substantially more risk than the periods before and after. And even if transition risks were higher then, there would be plenty of time and warning in the last few decades before such a transition to think carefully about how to design regimes of governance and control for large systems of interacting tools, before humans are displaced out of the main social roles of design, governance, and control.

The Continuation of Past Trends

Society has changed in many ways over the last century or two, and in this scenario society would continue to change at similar rates over the coming centuries before a transition to an economy dominated by machine intelligence. Since some of the changes in the last century or so have been enabled and driven by specific technologies, some future changes would also have this character.

However, while some future changes will no doubt be driven by particular technologies, it is important not to over-estimate the fraction of changes that arise this way. Most of the big changes we've seen over the last century or so have not been driven by particular technologies. Instead, most such changes can be better understood as resulting from general broad increases in wealth and capacity.

Fortunately, these sort of changes tend to be easier to predict. Increases in physical capacities suggest that we will continue to see increases in travel speeds and distances, in the size range of homes, vehicles, and meals, in lighting and brightness, in sound insulation and quietness, in the lightness, toughness, and colorfulness of clothing, in the climate control of artificial spaces, in building heights and depths, and in memory sizes and processing and communication speeds. And those later increases in info system capacities suggests we will know more about the actions and feelings of others, and have more decision aid tools.

Increases in social capacities suggest that that we will continue to see increases in lifespans, individual intelligence and abilities to navigate complex social systems, in the size, density, complexity, and specialization of firms and cities, in the specialization of individual roles and social networks. We also expect to see a steady increase in the quality and addictiveness of art, decorations, and entertainment.

Many trends over the last century or so can be understood as due to a reversion to forager values, habits, and attitudes, as wealth has weakened the strong social pressures that turned foragers into farmers. While at work we have come to accept increasing alienation and domination, such as via varying detailed orders and fine-grained status rankings, outside of work we have less tolerance for domination and overt ranking. In particular we are less tolerant of autocracy, slavery, and overt class distinctions. We should expect such trends to continue.

We also expect to seen a continued fall in fertility, and continued rise in leisure, peace, travel, promiscuity, romance, civility, mental-challenging work, local diversity of style, and medical and art spending. We expect to continue to see increasing value placed on self-direction, tolerance, pleasure, nature, leisure, and political participation.

A fast sudden transition

The other possible scenario that I'll consider in this chapter for a transition to a world dominated by machine intelligence is based on the arrival of the ability to make brain emulations, which I'll call "ems." An em would result from taking a particular human brain, scanning it to record its particular cell features and connections, and then building a computer model that processes signals according to those same features and connections.

A good enough em would have very close to the same overall input-output signal behavior as the original human. One might talk with it, and convince it to do useful jobs. When ems are cheap, they would quickly displace humans on almost all tasks.

Since an almost working emulation is of little use, while a working emulation can be very useful, this scenario can result in a more sudden and disruptive transition.

The ability to make ems mainly requires that three kinds of technology will sufficiently advance. Computer hardware must get cheap enough, brain scans must get cheap and fast enough with high enough spatial and chemical resolution, and computer models of specific types of brain cells must get accurate enough. Based on prior trends, these all seem likely to advanced sufficiently within about a century, or well before AI software is anywhere near human level abilities (Sandberg & Bostrom 2008).

We can foresee many ways in which a world dominated by ems would change. But such changes are mostly beyond the scope of this book, which is mainly concerned with the transition to such a world. However, to understand this transition, it helps to understand some basics about this new world.

In particular, it helps to understand that net income in this new world would go to those who own relevant land and natural resources, to those who own and run the firms who fill key industry niches, and to the ems who would dominate most labor market niches, and to those who own relevant capital, not only machines and buildings but also the computer hardware that runs ems.

The transition to an em world would result in big changes in who gets most world income. The em-based economy would quickly come to dominate the world economy. Ems would probably concentrate in a few dense new cities. The switch to em-based capital and computer-hardware would likely put a new crop of firms in key industries niches. And a few hundred “clans” of ems all descended from the same original human would probably dominate most em labor markets.

First-movers in these areas will tend to be disproportionate winners. These include the first firms to make and distribute em hardware, the real estate areas that house the first em activities, the first humans to be scanned to become em workers, and those who hold patents on key enabling technologies. In anticipation of this big disruption of income, many would try to scramble to position themselves to be transition winners. Those who find themselves falling behind in this race may be tempted to resort to violence.

In the AI software based scenario discussed above, the transition is expected to happen later, to be slower and smoother, and to give more warning of a transition ahead. Given sufficient advanced preparation, that transition needn’t be especially disruptive, since there is less reason to expect big changes in the value of patents, real estate, or key firms during such a transition.

In contrast, a brain emulation based scenario can have a more disruptive transition, because having a working em is more of an all or nothing situation. While tools can slowly get more useful as they get more capable, an em mostly either works or it doesn't. Thus there can be much less warning about an upcoming em transition, which may happen sooner, more quickly, and make for bigger changes between winners and losers. The transition to an em world can be especially disruptive and unexpected if the last technology to be ready is cell modeling or brain scanning.

These differences would make it more important for individuals and their supporting institutions to prepare more carefully for a possible transition, even in the absence of clear warnings of an imminent transition. Assets should be diversified and insured more thoroughly, and based on weaker signs regarding upcoming problems. These differences also make it more likely that such preparations will not be made, so that big winners and losers will result.

Well before a transition to a world dominated by em-based machine intelligences, the world would much like it would well before a transition to a world dominated by human level AI software. In both cases, our physical and social capacities would be increasing, and our values would be slowly becoming more forager-like.

However, while our physical and social capacities would continue to increase after either transition, value trends would more clearly change after an em transition. Soon the vast majority of creatures with human like values would be ems, and the fall of wages to near em subsistence levels would likely induce a sudden and large reversal of our industry-era trend toward forager-like values and attitudes. The anticipation of this change would likely also add to the stress and disruption of a transition to an em-dominated world.

Conclusion

In this chapter I have considered and compared two quite different scenarios for a future transition to a world dominated by machine intelligence. One scenario is based on continuing on our current path of accumulating better software, and results in the smooth continuation of current trends for centuries, until a relatively gradual and anticipated transition, when there are relatively mild disruptions to power and values. The other scenario is based on the future arrival of an ability to fully emulate human brains, and is by its nature sooner, more sudden, and more disruptive to both power and values.

In my judgment this second of these two scenarios is the more likely one. Which is why I have been working on a detailed book-length analysis of its likely post-transition consequences. Even if it is not the transition or world we would most wish to have, if it is the world we will have, it is important to understand in as much detail as possible, to inform any decisions we may make about it.

Daron Acemoglu (2009) *Introduction to Modern Economic Growth*, Princeton University Press, January.

Stuart Armstrong, Kaj Sotala (2012) How we're predicting AI-or failing to. In J. Romportl, P. Ircing, E. Zackova, M. Polak, and R. Schuster, R. editors, *Beyond AI: Artificial Dreams*, 52-75, Pilsen: University of West Bohemia.

Hadi Esmaeilzadeh, Emily Blem, Renee Amant, Karthikeyan Sankaralingam (2012) Power Limitations and Dark Silicon Challenge the Future of Multicore. *Transactions on Computer Systems* 30(3):11 August.

Katja Grace (2013) Algorithmic Progress in Six Domains, Technical report 2013-3, Machine Intelligence Research Institute, October 5. http://intelligence.org/files/AlgorithmicProgress.pdf

Robin Hanson (2012) AI Progress Estimate, *Overcoming Bias* blog, August 27, http://www.overcomingbias.com/2012/08/ai-progress-estimate.html .

Robin Hanson, Eliezer Yudkowsky (2013) *The Hanson-Yudkowsky AI-Foom Debate*, Berkeley, CA: Machine Intelligence Research Institute.

Loukas Karabarbounis, Brent Neiman (2013) The Global Decline of the Labor Share, NBER Working Paper No. 19136, June.

Devendra Sahal (1981) *Patterns of technological innovation*. Addison-Wesley.

Anders Sandberg & Nick Bostrom (2008) Whole Brain Emulation: A Roadmap, Technical Report #2008-3, Future of Humanity Institute, Oxford University. http://www.fhi.ox.ac.uk/__data/assets/pdf_file/0019/3853/brain-emulation-roadmap-report.pdf