# Three impacts of machine intelligence

Paul Christiano Follow
Nov 29, 2014 · 10 min read

I think that the development of human level AI in my lifetime is quite plausible; I would give it more than a 1-in-4 chance. In this post I want to briefly discuss what I see as the most important impacts of AI. I think these impacts are the heavy hitters by a solid margin; each of them seems like a big deal, and I think there is a big gap to #4.

- **Growth will accelerate, probably very significantly.** Growth rates will likely rise by at least an order of magnitude, and probably further, until we run into severe resource constraints. Just as the last 200 years have experienced more change than 10,000 BCE to 0 BCE, we are likely to see periods of 4 years in the future that experience more change than the last 200.
- **Human wages will fall, probably very far.** When humans work, they will probably be improving other humans' lives (for example, in domains where we intrinsically value service by humans) rather than by contributing to overall economic productivity. The great majority of humans will probably not work. Hopefully humans will remain relatively rich in absolute terms.
- **Human values won't be the only thing shaping the future.** Today humans trying to influence the future are the only goal-oriented process shaping the trajectory of society. Automating decision-making provides the most serious opportunity yet for that to change. It may be the case that machines make decisions in service of human interests, that machines share human values, or that machines have other worthwhile values. But it may also be that machines use their influence to push society in directions we find uninteresting or less valuable.

My guess is that the first two impacts are relatively likely, that there is unlikely to be a strong enough regulatory response to prevent them, and that their net effects on human welfare will be significant and positive. The third impact is more speculative, probably negative, more likely to be prevented by coordination (whether political regulation, coordination by researchers, or something else), and also I think more important on a long-run humanitarian perspective.

None of these changes are likely to occur in discrete jumps. Growth has been accelerating for a long time. Human wages have stayed high for most of history, but I expect them to begin to fall (probably unequally) long before everyone becomes unemployable. Today we can already see the potential for firms to control resources and make goal-oriented decisions in a way that no individual human would, and I expect this potential to increase continuously with increasing automation.

Most of this discussion not particularly new. The first two ideas feature prominently in Robin Hanson's speculation about an economy of human emulations (alongside many other claims); many of the points below I picked up from Carl Shulman; most of them are much older. I'm writing this post here because I want to collect these thoughts in one place, and I want to facilitate discussions that separate these impacts from each other and analyze them in a more meaningful way.

# Growth will accelerate

There are a number of reasons to suspect that automation will eventually lead to much faster growth. By "much faster growth" I mean growth, and especially intellectual progress, which is at least an order of magnitude faster than in the world of today.

I think that avoiding fast growth would involve solving an unprecedented coordination problem, and would involve large welfare losses for living people. I think this is very unlikely (compare to environmental issues today, which seem to have a lower bar for coordination, smaller welfare costs to avert, and clearer harms).

## Automating tech progress leads to fast growth.

The stereotyped story goes: "If algorithms+hardware to accomplish X get 50% cheaper with each year of human effort, then they'll also (eventually) get 50% cheaper with each year of AI effort. But then it will only take 6 months to get another 50% cheaper, 3 months to get another 50% cheaper, and by the end of the year the rate of progress will be infinite."

In reality things are very unlikely to be so simple, but the basic conclusion seems quite plausible. It also lines up with the predictions of naive economic models, on which constant returns to scale (with fixed tech) + endogenously driven technology—> infinite returns in finite time.

Of course the story breaks down as you run into diminishing returns to intellectual effort, and once “cheap” and “fast” diverge. But based on what we know now it looks like this breakdown should only occur very far past human level (this could be the subject for a post of its own, but it looks like a pretty solid prediction). So my money would be on a period of fast progress which ends only once society looks unrecognizably different.

One complaint with this picture is that technology already facilitates more tech progress, so we should be seeing this process underway already. But we do see accelerating growth (see the section on the historical record, below), except for a historically brief period of 50–75 years. So this seems like a weak objection.

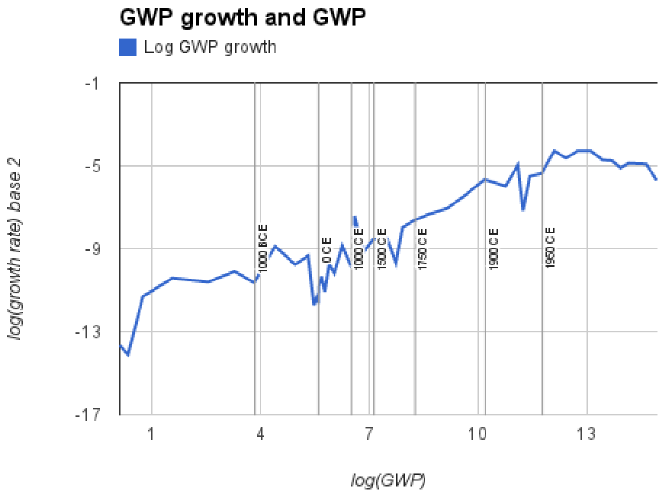
## Substituting capital for labor leads to fast growth.

Even if we hold fixed the level of technology, automating human labor would lead to a decoupling of economic growth from human reproduction. Society could instead grow at the rate at which robots can be used to produce more robots, which seems to be much higher than the rate at which the human population grows, until we run into resource constraints (which would be substantially reduced by a reduced dependence on the biosphere).

## Extrapolating the historical record suggests fast growth.

Over the course of history the proportional rate of growth has increased substantially, from more than 1,500 years per doubling around 10k years ago, to around 150 years per doubling in the 17th century, to around 15 years per doubling in the 20th century. A reasonable extrapolation of the pre-1950 data appears to suggest an asymptote with infinite population sometime in the 21st century. The last 50 years represent a notable departure from this trend, although history has seen similar periods of relative stagnation.

(See the graph to the left, produced from Bradford Delong's data here; data and full size here. The graph starts at a doubling time of 8000 years, and plateaus around 25 years, about 300 times faster. The average time required for growth rates to double is about 1.5 doublings, or 40 years at the current rate. Each subsequent doubling of the growth rate takes half as long.)



I don't think that extrapolating this trend forward results in particularly robust or even meaningful predictions, but I do think it says one thing: we shouldn't be surprised by a future with much faster growth. The knee jerk response of "that seems weird" is inconsistent with the actual history (though it is a very reasonable intuitive reaction for someone who lived through the 2nd half of the 20th century). The more recent trend of slow growth may well continue, but I don't think we should be surprised if this stagnation was a temporary departure from trend, comparable to previous departures in severity and scale.

# Wages will fall

I think this is the most robust of the three predictions. It seems very likely that eventually machines will be able to do all of the things that a human can do, as well as a human can do it. At that point, it would require significant coordination to prevent the broad adoption of machines as human replacements. Moreover, continuing to use human labor in this scenario seems socially undesirable; if we don't have to work it seems crazy to make work for ourselves, and failing to make use of machine labor would involve even more significant sacrifices in welfare.

(Humans may still demand other humans' labor in distinctively human roles. And there may be other reasons for money to move around between humans. But overall most new valuable stuff and valuable ideas in the world will be produced by machines. In the optimistic scenario, the main reason that this value will be flowing *to* humans, rather than merely *amongst* humans, will be because they own some of the machines.)

Historically humans have not been displaced by automation. I think this provides some evidence that in the near term automation will not displace humans, but in the long run it looks inevitable, since eventually machines really will be better at *everything*. Simple theories do suggest a regime where humans and automation are complementary, followed by a regime in which they are substitutes (as horses were once complementary with carriages, but were eventually completely replaced by automation). Robin Hanson illustrates in slides 30–32 here. So at some point I think we should expect a more substantive transition. The most likely path today seems to be a fall in wages for many classes of workers while driving up wages for those who are still complementary with automation (a group that will shrink until it is eventually empty). "Humans need not apply" has been making the rounds recently; and despite many quibbles I think it does a good job of making the point.

I don't have too much to say on this point. I should emphasize that this isn't a prediction about what will happen soon, just about what will have happened by the time that AI can actually do everything humans can do. I'm not aware of many serious objections to this (less interesting) claim.

# Human values won't be the only things shaping the future

I'd like to explicitly flag this section as more speculative. I think AI opens up a new possibility, and that this is of particular interest for those concerned with the long-term trajectory of society. But unlike the other sections, this one rests on pretty speculative abstractions.

Many processes influence what the world will look like in 50 years. Most of those processes are not goal-directed, and push things in a somewhat random direction; an asteroid might hit earth and kill us all, the tectonic plates will shift, we'll have more bottles in landfills because we'll keep throwing bottles away, we'll send some more radiation into space. One force stands out amongst all of these by systematically pushing in a particular direction: humans have desires for what the future looks like (amongst other desires), and they (sometimes) take actions to achieve desired outcomes. People want themselves and their children to be prosperous, and so they do whatever they think will achieve that. If people want a particular building to keep standing, they do whatever they think will keep it standing. As human capacities increase, these goal-oriented actions have tended to become more important compared to other processes, and I expect this trend to continue.

There are very few goal-directed forces shaping the future aside from human preferences. I think this is a great cause for optimism: if humans survive for the long run, I expect the world to look basically how we want it to look. As a human, I'm happy about that. I don't mean "human preferences" in a narrow way; I think humans have a preference for a rich and diverse future, that we care about other life that exists or could have existed, and so on. But it's easy to imagine processes pushing the world in directions that we don't like, like the self-replicator that just wants to fill the universe with copies of itself.

We can see the feeble beginnings of competing forces in various human organizations. We can imagine an organization which pursues its goals for the future even if there are no humans who share those goals. We can imagine such an organization eventually shaping what other organizations and people exist, campaigning to change the law, developing new technologies, *etc.*, to make a world better suited to achieving its goals. At the moment this process is relatively well contained. It would be surprising (though not unthinkable) to find ourselves in a future where 30% of resources were controlled by PepsiCo, fulfilling a corporate mission which was completely uninteresting to humans. Instead PepsiCo remains at the mercy of human interests, by design and by necessity of its structure. After all, PepsiCo is just a bunch of humans working together, legally bound to behave in the interest of some other humans.

As automation improves the situation may change. Enterprises might be autonomously managed, pursuing values which are instrumentally useful to society but which we find intrinsically worthless (e.g. PepsiCo can create value for society even if its goal is merely maximizing its own profits). Perhaps society would ensure that PepsiCo's values are to maximize profit only insofar as such profitmaximization is in the interest of humans. But that sounds complicated, and I wouldn't make a confident prediction one way or the other. (I'm talking about firms because its an example we can already see around us, but I don't mean to be down on firms or to suggest that automation will be most important in the context of firms.)

At the same time, it becomes increasingly difficult for humans to directly control what happens in a world where nearly all productive work, including management, investment, and the design of new machines, is being done by machines. We can imagine a scenario in which humans continue to make all goal-oriented decisions about the management of PepsiCo but are assisted by an increasingly elaborate network of prosthetics and assistants. But I think human management becomes increasingly implausible as the size of the world grows (imagine a minority of 7 billion humans trying to manage the equivalent of 7 trillion knowledge workers; then imagine 70 trillion), and as machines' abilities to plan and decide outstrip humans' by a widening margin. In this world, the AI's that are left to do their own thing outnumber and outperform those which remain under close management of humans.

Moreover, I think most people don't much care about whether resources are held by agents who share their long-term values, or machines with relatively alien values, and won't do very much to prevent the emergence of autonomous interests with alien values. On top of that, I think that machine intelligences can make a plausible case that they deserve equal moral standing, that machines will be able to argue persuasively along these lines, and that an increasingly cosmopolitan society will be hesitant about taking drastic anti-machine measures (whether to prevent machines from having "anti-social" values, or reclaiming resources or denying rights to machines with such values).

Again, it would be possible to imagine a regulatory response to avoid this outcome. In this case the welfare losses of regulation would be much smaller than in either of the last two, and on a certain moral view the costs of no regulation might be much larger. So in addition to resting on more speculative abstractions, I think that this consequence is the one most likely to be subverted by coordination. It's also the one that I feel most strongly about. I look forward to a world where no humans have to work, and I'm excited to see a radical speedup in technological progress. But I would be sad to see a future where our descendants were maximizing some uninteresting values we happened to give them because they were easily specified and instrumentally useful at the time.

· · ·

*Originally published at rationalaltruist.com on August 23, 2014.*