AI Impacts

Katja Grace Paul Christiano

May 17, 2015

1. Introduction

AI Impacts answers policy-relevant questions about the long-term future of AI. We do this by conducting research projects in a range of subject areas. Our research interests are well represented by Section 4 of FLI's 'A survey of research questions for robust and beneficial AI'.[1] We emphasize long-term issues, such as how soon AI will be capable of replacing humans in most jobs, how this will affect society, and how quickly these effects will take place. However our research naturally bears on related short-term issues, such as the economic impacts of automation, and near-term trends in hardware and AI progress.

At a high level, our priority questions are:

1. What should we believe about timelines for AI development?
2. How rapid is the development of AI likely to be near human-level? How much advance notice should we expect to have of disruptive change?
3. What are the likely economic impacts of human-level AI?
4. Which paths to AI should be considered plausible or likely?
5. Will human-level AI tend to pursue particular goals, and if so what kinds of goals?
6. Can we say anything meaningful about the impact of contemporary choices on long-term outcomes?

However, our research is presently on lower level questions, which inform these.

We publish our research articles at aiimpacts.org, and describe our processes and findings informally on our research blog (http://aiimpacts.org/blog/). All work on AI Impacts is in the public domain.

AI Impacts exists because we perceive a lack of high quality empirical and integrative research informing policy and other strategic decisions on the future of AI. This project is designed to remedy that, and also to provide good organization and exposition of existing research on the topic.

The project is organized as a collection of small research projects, which use a variety of methods as appropriate. These are mostly the methods of social science. For instance, in recent times we have collected data from primary and secondary sources, analyzed it statistically, interviewed experts, investigated case studies, and applied previous findings in fields like history, economics and genetics to questions about AI.

Overall, we aim to answer forecasting questions that are both subject to tractable methods, and particularly informative to people seeking socially beneficial outcomes from AI. We hope to inform policy analyses, legal analyses, funding bodies' consideration of differential technological progress, and other investment decisions. We also hope to inform educated opinion on AI and its social implications. We write for an

[1] http://futureoflife.org/static/data/documents/research_survey.pdf

audience of AI researchers, AI risk researchers, policymakers, philanthropists, educated laypeople, and academic researchers in related subjects.

2. Preliminary work

Katja has worked at the Machine Intelligence Research Institute (MIRI) for a year and a half, largely on empirical research closely related to that at AI Impacts. This section will briefly describe some of this early work. The next section will discuss in more detail the work that we propose.

The characteristic speed and nature of progress on algorithmic problems is evidence about how AI capabilities will grow in the coming decades. Katja conducted a preliminary investigation into algorithmic progress in 2013, which became *Algorithmic progress in six domains*.[2] She collected and analyzed data on progress in Boolean satisfiability, chess, Go, factoring, various machine learning applications, scheduling, linear optimization, and physics simulations. Across these areas, improvements over time from algorithmic progress tended to be relatively smooth. Algorithmic improvements contributed around half as much to progress as hardware improvements, for the problem sizes being used.

An understanding of generic progress in computer science is important because it informs our expectations about timing of AI, as well as our expectations about how suddenly or incrementally human-level AI is likely to appear. We hope to further investigate progress in algorithms as part of AI Impacts. Natural next steps from this project include investigating progress on theoretical algorithms, and investigating progress on algorithmic problems selected for being economically important or close to AI rather than for having available data.

While at MIRI, Katja has also analyzed two historical efforts to defend humanity from technological risks far ahead of time. These were Leo Szilard's attempts to make information about nuclear weapons secret before and during WWII, and efforts to avoid man-made pandemics around the Asilomar Conference on Recombinant DNA[3]. A primary goal of this research was to learn whether attempts to prepare for risks decades ahead of time tend to be worthwhile, and especially whether attempts in cases like AI risk tend to be worthwhile.

Some effort went into assessing the similarity of these cases to that of AI risk, and their success. The project also sought other concrete implications for contemporary AI risk efforts. For instance, what factors make a difference to success? Is it important to have experts behind the cause? How does public fear affect outcomes? A big part of these projects was interviewing experts. Katja talked to two organizers of the Asilomar

[2] https://intelligence.org/files/AlgorithmicProgress.pdf
[3] Both forthcoming. They should be available at the top of the MIRI 'All Publications' page by the end of May (https://intelligence.org/all-publications/)

Conference, two experts on nuclear history, and a person from GiveWell who believed that early risk mitigation is low value.[4]

The algorithmic progress, Szilard and Asilomar projects are all similar to what we hope to do with AI Impacts, though AI Impacts is organized as smaller, modular investigations. We also have preliminary work on AI Impacts, but will discuss this in later sections, since it is incomplete and we are proposing to continue with it.

3. Proposed work

We propose to conduct a range of small research projects designed to efficiently inform anticipations of strong artificial intelligence. We hope to expand several lines of research that we have already made progress on. We describe two of these below (sections 3.1-3.2). We also hope to begin on several new research projects (outlined in section 3.3). Some of these we have preliminary work on, but are yet to publish it.

3.1 Measuring the brain in TEPS

We recently made a preliminary estimate of the computing power of the brain as measured in traversed edges per second (TEPS), a measure of communications performance usually used for supercomputers.[5] We detail it here both as a line of research we would like to work more on, and as an example of the type of research we might do on different AI forecasting topics. This section borrows parts from our pages on this topic, especially 'A new approach to predicting brain-computer parity'.[6]

We are interested in comparing the brain's capabilities to that of contemporary computers, and those anticipated in coming years. Such comparisons inform expectations about when AI will become broadly human-level (though not straightforwardly—we are also interested in what human-equivalence in hardware would imply about AI, and have written some about it)[7].

Various estimates have been made of the brain's computing ability, in terms of performance benchmarks such as floating point operations per second (FLOPS) and millions of instructions per second (MIPS).[8] These measure how fast a computer can perform individual operations. However a computer also needs to move information around between the various components performing operations. This communication

---

[4] Notes from these conversations are forthcoming, by May 31, 2015.
[5] http://aiimpacts.org/brain-performance-in-teps/
[6] http://aiimpacts.org/tepsbrainestimate/
[7] http://aiimpacts.org/how-ai-timelines-are-estimated/
[8] Moravec, 2009, http://www.scientificamerican.com/article/rise-of-the-robots/
Moravec, 1997, http://www.transhumanist.com/volume1/moravec.htm
Kurzweil, 2005, The Singularity is Near, Viking
Sandberg & Bostrom, 2008, http://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf

takes time, space and material, and so can substantially affect overall performance of a computer, especially on data-intensive applications. Consequently when comparing computers it is useful to have performance metrics that emphasize communication as well as ones that emphasize computation. We argue that this is particularly likely to be relevant when measuring brains, which appear to be powerful enough to be bottlenecked on communication and also appear to use substantial resources on communication.[9]

If communication is a bottleneck, then it is especially important to know when computers will achieve similar performance to the brain there, not just on easier aspects of being a successful computer. So this is is one reason to measure the brain in terms of TEPS. It is also good to have estimates of the brain's power based on relatively independent measures, since all are fairly uncertain. Furthermore an analog to TEPS in the brain is relatively easy to measure.

The TEPS benchmark asks the computer to simulate a graph, and then to search through it.[10] The question is how many edges in the graph the computer can follow per second. The brain can't run the TEPS benchmark. However the brain is itself a graph of neurons, and we can measure edges being 'traversed' in it: action potentials communicating between neurons. So we can count how many edges are traversed in the brain per second, and compare this to existing computer hardware.

A review of the literature suggests that the brain has around 1.8-3.2 x 10^{14} synapses.[11] We'd like to know how often these synapses convey spikes, but have not found good data on it. So we use neuron firing frequency as a proxy. We estimate that each neuron spikes around 0.1-2 times per second, based on several lines of experimental evidence.[12] Together with the number of synapses, this suggests the brain performs at around 0.18 – 6.4 * 10^{14} TEPS. This is somewhere between as powerful and thirty times as powerful as the best supercomputer, in terms of TEPS. Our estimate relies on many assumptions, which we discuss elsewhere.[13] The estimate could be improved on many fronts with more work.

We also calculated current prices for TEPS, based on publicly cited prices for supercomputers for which TEPS measurements are available (all from the Graph 500 list).[14] We estimate that a gigaTEPS can be purchased for around $0.26/hour. Given this price, performance equivalent to that of the brain in terms of TEPS should cost roughly $4,700 – $170,000/hour. Our best guess is that TEPS prices will improve by a factor of ten every four years, largely because other benchmarks improve at that rate.[15] If

[9] http://aiimpacts.org/brain-performance-in-teps/
[10] http://www.graph500.org/specifications
[11] http://aiimpacts.org/scale-of-the-human-brain/
[12] http://aiimpacts.org/rate-of-neuron-firing/
[13] http://aiimpacts.org/brain-performance-in-teps/
[14] http://aiimpacts.org/cost-of-teps/
[15] http://aiimpacts.org/cost-of-teps/

this is true, it should take seven to fourteen years for a computer which costs $100/hour to be competitive with the human brain, in terms of TEPS.

This evidence points to human-level hardware being available in around a decade, which is quite close to Kurzweil's estimate based on computation (four years), and Sandberg and Bostrom's more optimistic estimate for when hardware will exist to emulate a brain (twelve years).[16] Moravec predicts earlier, and Sandberg and Bostrom predict later if deeper levels of the brain need to be represented.

Our estimate so far is preliminary. We would like to improve it by strengthening many instrumental findings. For instance, we would like to better estimate the rate of neural firing, and the number of synapses. We would also like to have figures for average rates of firing per synapse. Better measurements of the rate at which TEPS become cheaper over time would be helpful, as would data on the cost of TEPS in computers other than supercomputers.

The estimate could also be improved by checking the validity of various assumptions. For instance, we have assumed that the information contained in neural spikes is similar to that transmitted in the TEPS benchmark, and that the exact distribution of links in the graph doesn't matter a lot. These issues could be investigated with reference to empirical evidence.

Our estimate is somewhat complex, and draws on evidence from different disciplines. We plan to talk with experts from those disciplines about what we have done, to check that our analysis is valid given further facts about neuroscience, hardware progress, or communications benchmarking. We also hope they might have further ideas for how to estimate difficult quantities within their area of expertise (e.g. what resources are spent on communication in the brain; how fast TEPS performance is improving). We have already discussed our preliminary TEPS estimate with two people with neuroscience expertise, who broadly agreed with our methods and gave us useful feedback and pointers to relevant research. We plan to publish notes on some such discussions in the future.

We may also pursue some of several closely related lines of inquiry, which are not part of this project, but complement it well. We would like to think more about the implications of affordable human-level hardware being soon. This will likely involve discussion with thinkers on the topic. We would also like to investigate the relationship in general between hardware, software and AI. This might be done via a variety of research projects.[17] For instance, we could try to separate the contributions of hardware and software progress to specific contemporary AI progress for which we have data (as Katja did sometimes in *Algorithmic Progress in Six Domains*). Another valuable related project is to check, update, and expand on older measures of the brain's computing power in terms of FLOPS and other computing metrics.

---

[16] http://aiimpacts.org/preliminary-prices-for-human-level-hardware/

[17] We have a list of research ideas relating to this which we plan to publish soon. It should be linked from http://aiimpacts.org/possible-investigations/ when it is available.

3. 2 Characterizing abrupt progress

We are interested in whether AI research might undergo discontinuous progress in the lead-up to human-level capabilities, or whether progress will be smooth. We are also interested in the scale of any discontinuities. To learn about these issues, we are collecting case studies of discontinuous progress in technology.[18] This should be part of a larger investigation into the probability of abrupt progress. This section borrows passages from some of our pages.[19]

There are two kinds of reasons we are interested in how abruptly AI progress may proceed: anticipating abrupt progress would change our predictions, and also change what kind of forecasting research is applicable. If discontinuity is likely, a transition to AI is more likely to be abrupt, more likely to be soon, and more likely to be disruptive. Also, if we think a discontinuity is likely, then our research should investigate questions such as how to foresee or mitigate the effects of discontinuities, and not questions like when the present trajectories of AI progress will reach human-level capabilities. As well as being decision relevant and important, this question appears to attract substantial disagreement, making it particularly important to resolve.

This project aims to shed light on the potential for discontinuities in AI by investigating the degree and nature of discontinuities in other technologies. This seems an informative baseline for our expectations about AI, especially if we have no strong reason to expect artificial intelligence to be radically unusual in this regard. It will also allows us to test theories about what features of AI might make it more or less likely to undergo abrupt progress. By considering other technologies, we can evaluate whether more virtual technologies do tend to see sudden progress often, and whether the scale of any such discontinuities is large enough to change our expectations about AI timelines.

We have collected around fifty instances of technological change that are contenders for being discontinuous, mostly suggested to us by others. We are learning about these cases one by one, and assessing whether each involved discontinuous progress on any interesting metrics. So far we have made detailed assessments of six potentially discontinuous technologies, and preliminary assessments of another six.[20]

We have been measuring the scale of discontinuous progress in terms of ‘years of usual progress in a single event’. So far, we have found two instances of more than one hundred years of usual progress in a single event. Nuclear weapons marked around six thousand years of progress in ‘relative effectiveness’ of explosives (explosive power per weight). High temperature superconductors marked more than one hundred years of

[18] http://aiimpacts.org/discontinuous-progress-investigation/
[19] http://aiimpacts.org/discontinuous-progress-investigation/
[20] http://aiimpacts.org/discontinuous-progress-investigation/

progress in the maximum temperature of superconductors. We found two other instances of smaller abrupt progress.[21]

Because nuclear weapons represented such a large discontinuity, we investigated the case more thoroughly.[22] We wanted to learn more about what made it unusual, and to use it as evidence on prevailing theories about what causes progress to generally be incremental. We roughly evaluated the cost-effectiveness of explosives over time, and did not find abrupt progress in that metric.[23] We informally analyzed several alternative explanations for nuclear weapons being unusual.[24] We explained this research in a series of blog posts intended for an educated lay audience (as well as in more formal online articles).[25]

We plan to continue this project by evaluating more cases of potentially abrupt technological progress. We plan to make these case studies part of a larger investigation into the likelihood of abrupt progress in AI. This would involve a deeper survey of the literature on technological progress, and discussion with experts on technological history, AI and specific case studies. We would also like to talk more with people who argue that progress is likely to be abrupt or not, to collect the arguments people find compelling, which we might then be able to evaluate. In this way, we hope to build an educated estimate of the plausibility of abrupt progress in AI prior to AI becoming competitive with humans.

This kind of research is also complementary with the measurement of algorithmic progress mentioned earlier, because the nature of progress in technological fields closely related to AI is especially informative.

3.3 Further potential projects

These are the main clusters of articles we have so far, all of which we plan to continue expanding:

1. **The brain measured in TEPS:** see section 3.1 for details
2. **Abrupt progress case studies:** see section 3.2 for details
3. **The range of 'human-level' intelligence**: how much research effort will it take to move from 'human-level' AI to AI that is soundly superior to any human?[26] This is important for predicting the disruptiveness of a transition to an AI-based economy.

---

[21] http://aiimpacts.org/cases-of-discontinuous-technological-progress/
[22] http://aiimpacts.org/discontinuity-from-nuclear-weapons/
[23] http://aiimpacts.org/were-nuclear-weapons-cost-effective-explosives/
[24] http://aiimpacts.org/whats-up-with-nuclear-weapons/
[25] http://aiimpacts.org/the-biggest-technological-leaps/
http://aiimpacts.org/ai-and-the-big-nuclear-discontinuity/
http://aiimpacts.org/whats-up-with-nuclear-weapons/
http://aiimpacts.org/were-nuclear-weapons-cost-effective-explosives/
[26] http://aiimpacts.org/the-slow-traversal-of-human-level/

4. **Relevant facts about the brain:** figures such as the number of neurons in the brain, and how frequently they fire.[27] These things are relevant to many other projects, such as measuring the brain in TEPS.
5. **Hardware trends**: how the price and quantity of computing hardware changes over time, and is predicted to change in the future.[28] We have investigated this in aid of measuring the brain in TEPS, but these figures are widely applicable.

These are clusters of supplementary reference pages that we have begun and plan to expand:

1. **Analyses of time to AI**: summaries of reasoning others have presented to estimate when various forms of strong AI will arrive.[29]
2. **Surveys**: summaries of expert surveys on AI timelines.[30]
3. **Terminology**: we have a small amount of discussion of 'human-level AI', and plan to clarify other terms used in the AI safety field.[31]
4. **Research ideas**: we collect research ideas broadly appropriate for AI impacts. We are in the process of writing short summaries of them, and making quick evaluations of their cost-effectiveness.[32]

These are clusters we anticipate beginning or continuing from a modest (unpublished) start:

1. **Analyses of MIRI data on public AI predictions:** MIRI previously collected a large dataset of public AI predictions. We have improved it somewhat, and are analyzing its implications regarding timing as well as bias and accuracy.
2. **Review of technology forecasting track record:** a basic input to evaluating predictions of AI timelines, and making more, is understanding what kinds of predictions are accurate, and how accurate they tend to be. We plan to review this at more length and summarize our findings.
3. **The computational cost of rerunning evolution to create human-level intelligence**: this is an upper bound on resources required to create human-level

---

[27] e.g. http://aiimpacts.org/scale-of-the-human-brain/ and http://aiimpacts.org/rate-of-neuron-firing/

[28] e.g. http://aiimpacts.org/current-flops-prices/, http://aiimpacts.org/cost-of-teps/, and http://aiimpacts.org/trends-in-the-cost-of-computing/.

[29] We list those we know about at http://aiimpacts.org/list-of-analyses-of-time-to-human-level-ai/ and detail two at http://aiimpacts.org/kurzweil-the-singularity-is-near/ and http://aiimpacts.org/allen-the-singularity-isnt-near/.

[30] This page links to nine more detailed summaries: http://aiimpacts.org/ai-timeline-surveys/.

[31] http://aiimpacts.org/human-level-ai/ and http://aiimpacts.org/at-least-human-level-at-human-cost-ai/

[32] http://aiimpacts.org/possible-investigations/ links to other lists of ideas. http://aiimpacts.org/research-topic-hardware-software-and-ai/ begins an unfinished series of more formal and evaluated research suggestions.

AI, and also bears on the difficulty of the problem in general. Bostrom has an estimate, which we are interested in building upon.[33]

4. **Models of intelligence explosion dynamics**: an 'intelligence explosion' is a hypothesized feedback loop between AI development, and AI labor available for further development. Various economic models have been suggested for it. Such an event might be very disruptive, so it would be valuable to better evaluate its plausibility, likely speed, and the conditions under which it is likely. This may involve collaboration with an economist.
5. **Disentangling contributions from hardware and software in contemporary AI progress**: this would inform our understanding of whether human-level AI was likely to arrive at around the time human-level hardware is affordable, or whether software is more important. Substantial evidence on this might both change expected AI timelines substantially, and also change what research we should do to predict them. Several kinds of investigations that might bear on this question.[34] We think this is a high priority.
6. **Interviews with experts:** expertise from AI researchers and other relevant experts would improve many of our projects. Deeper discussion with researchers would also greatly help with interpreting expert survey data on AI timelines. We have already interviewed four people working in AI or neuroscience, and hope to publish notes on these conversations soon. We plan to show completed sections of research to relevant experts who may disagree, for feedback on the most contentious parts.
7. **Relevance of neuroscience to AI progress**: neuroscience is often considered a plausible source for the software required to make human-level AI. We are interested in assessing this claim. We have begun this project by interviewing people working in AI and neuroscience on the connection between the two.
8. **Measuring progress in neuroscience:** if neuroscience might bear strongly on AI progress, then it matters how fast neuroscience is progressing. This is unclear. The rate of progress in neuroscience appears to be the disagreement at the heart of Kurzweil and Allen's different views on AI timelines[35]. We have talked to an expert about how to approach such measurement, and have begun to characterize progress in microscopy.
9. **Resolution of mathematical conjectures**: we outsourced the collection of dates that historical mathematical conjectures were posed and resolved. This might provide a prior distribution for how long to expect mathematical style problems to take, which bears on algorithmic progress in AI. We are yet to analyze the data and publish it.
10. **Historical economic growth trends:** because a transition to an AI economy would seem to be a large economic transition, we can draw some expectations about such an event from past large transitions, and economic history in general. For instance, we can ask how much and how quickly the economy has ever

[33] Bostrom, 2014, *Superintelligence*, Oxford (p25)
[34] List forthcoming, to appear on http://aiimpacts.org/possible-investigations/.
[35] http://aiimpacts.org/kurzweil-the-singularity-is-near/ and http://aiimpacts.org/allen-the-singularity-isnt-near/

changed before, as a result of various developments. This can inform expectations about the scale of future changes. Robin Hanson has written about these issues.[36] We have also done a small amount of unpublished analysis.

4. Expected outputs

In a year of work, we expect to have preliminary findings or substantially improved findings on most of the research projects we have already begun or anticipate beginning (see list in section 3.3). That is, around fifteen investigations, many of which have already received substantial work. For instance we expect to have a high quality estimate of the human brain's performance in TEPS. We hope to have a reasonable assessment of the plausibility of abrupt progress in AI research.

Each investigation corresponds to a cluster of smaller investigations and corresponding pages and blog posts. For example, we think of the brain in TEPS estimate as one larger investigation, and it includes around six articles and two blog posts. We also expect to opportunistically produce a number of supplementary reference pages, and minor investigations.

We plan to run a few small research workshops. We previously ran a successful afternoon workshop on research ideas relating to 'multipolar scenarios'. We would like to run similar events to seek counsel from relevant thinkers on other research questions.

5. Conclusion

To successfully navigate the advent of strong artificial intelligence, it is important to see where we are going. Serious AI forecasting has been neglected, and there are many tractable projects that could cost-effectively light our way. AI Impacts has made promising progress shedding light on decision-relevant considerations, and is well poised to further illuminate what lies ahead. These are perhaps pinpricks of light in a large and dimly lit space, but when there are only a few pinpricks, more illumination is often valuable.

[36] Hanson, 2000, http://mason.gmu.edu/~rhanson/longgrow.pdf