Monte Carlo model of brain emulation development

Future of Humanity Institute working paper 2014-1 (version 1.1) [1]

Anders Sandberg
Anders.sandberg@philosophy.ox.ac.uk
Future of Humanity Institute & Oxford Martin Programme on the Impacts of Future Technology
Oxford Martin School

Background

Whole brain emulation (WBE) is the possible future technology of one-to-one modelling of the function of the entire (human) brain. It would entail automatically scanning a brain, decoding the relevant neural circuitry, and generate a computer-runnable simulation that has a one-to-one relationship with the functions in the real brain (as well as an adequate virtual or real embodiment)[2].

Obviously this is a hugely ambitious project far outside current capabilities, possibly not even feasible in theory[3]. However, should such a technology ever become feasible there are good reasons to expect the consequences to be dramatic[4]: it would enable software intelligence, copyable human capital, new ethical problems, and (depending on philosophical outlook) immortality or a posthuman species. Even if one does not ascribe a high probability to WBE being ever feasible it makes sense to watch for trends indicating that it may be emerging, since adapting to its emergence may require significant early and global effort taking decades[5].

Predicting when a future technology emerges is hard, and there are good reasons to be cautious about overconfident pronouncements. In particular, predictions about the future of artificial intelligence have not been very successful and there are good theoretical reasons to have expected this[6]. However, getting a rough estimate of what is needed for a technology to be feasible compared to current trends can give a helpful "order of magnitude estimate" of how imminent a technology is, and how quickly it could move from a primitive state to a mature state.

This paper will describe a simple Monte Carlo simulation of the emergence of WBE as a first tool for thinking about it.

[1] Version history: 1.1 adds formulas and a more extensive description of the model, the requirements result section and some discussion about physics limits.
[2] (Sandberg & Bostrom 2008)
[3] (Sandberg 2014a)
[4] (Sandberg & Eckersley 2014)
[5] (Sandberg 2014b, Sandberg & Eckersley 2014)
[6] (Armstrong & Sotala 2013)

Method

The model is based on the following assumptions:

WBE will come about *at earliest* when

1. There is enough computing power to simulate the brain at a sufficiently high resolution.
2. There exists a project with enough budget to buy enough computing power.
3. There exists a scanning technology that can scan at least a single human brain at this. resolution.
4. There exists enough neuroscience understanding to turn the scan into a runnable model

Since these conditions cannot be predicted with a high certainty my model will treat them as (partially dependent) random variables in order to produce a probability distribution of the eventual arrival of WBE.

I am not making any particular assumptions about what technology is used to achieve necessary steps: the model only looks at overall abilities, not whether they are achieved through (for example) atomically precise manufacturing or exotic computing.

Necessary computing requirements

The computing power needed depends on the resolution where scale separation takes place: a more fine-grained simulation will not gain any better fidelity, while a coarser simulation will lack relevant functions. At present what resolution is needed is not known. The required resolution *R* is hence selected randomly, assuming a mean somewhere on the detailed electrophysiology side (based on the WBE workshop consensus) and a variance of one level[7].

$$R \sim N(5.5,1)$$

Given the resolution it is possible to estimate the number of entities (neurons, synapses, molecules etc. depending on resolution[8]) and hence the rough computational requirements *g(R)* for simulating each entity, producing a target computational requirement *C* for this scenario[9].

$$N_{entities} = f(R)$$

$$C = g(R) \cdot N_{entities}$$

Project size

The mode randomly generates how much money the computation is allowed to cost. This is between a million and a billion dollars, distributed as a truncated exponential distribution (i.e. uniformly distributed logarithm).

$$\log_{10} B \sim U(6,9)$$

[7] Levels defined in table 2 of the WBE report, (Sandberg & Bostrom 2008, p.13)
[8] (Sandberg & Bostrom 2008, pp. 79-81)
[9] The computational requirements used here are just the processing requirements, since storage requirements for WBE on a given level of resolution are very likely fulfilled many years before the corresponding processing requirements. A more elaborate model could separate storage and processing scenarios, but it would have to estimate their future correlation.

This is one area where the model can be refined by taking into account real research funding data and trends. It might also be beneficial to estimate budget constraints on scanning technology, which is currently not done in the model.

Computing power

Given a certain requirement and budget, when will there be enough computing power to run a WBE? This will depend on the available computer power in the future. Moore's law currently predicts an exponential increase in computing power per second and dollar, but long-term extrapolations must include a slowdown if only because of fundamental physical limits[10]. The model makes a fit of a sigmoidal function to resampled[11] data from Nordhaus's Moore's law data[12], producing a scenario for Moore's law with an eventual endpoint.

$$M(t) = c_1 + c_2\left[\frac{1}{2} + \frac{1}{2}\tanh(c_3(t - c_4))\right]$$

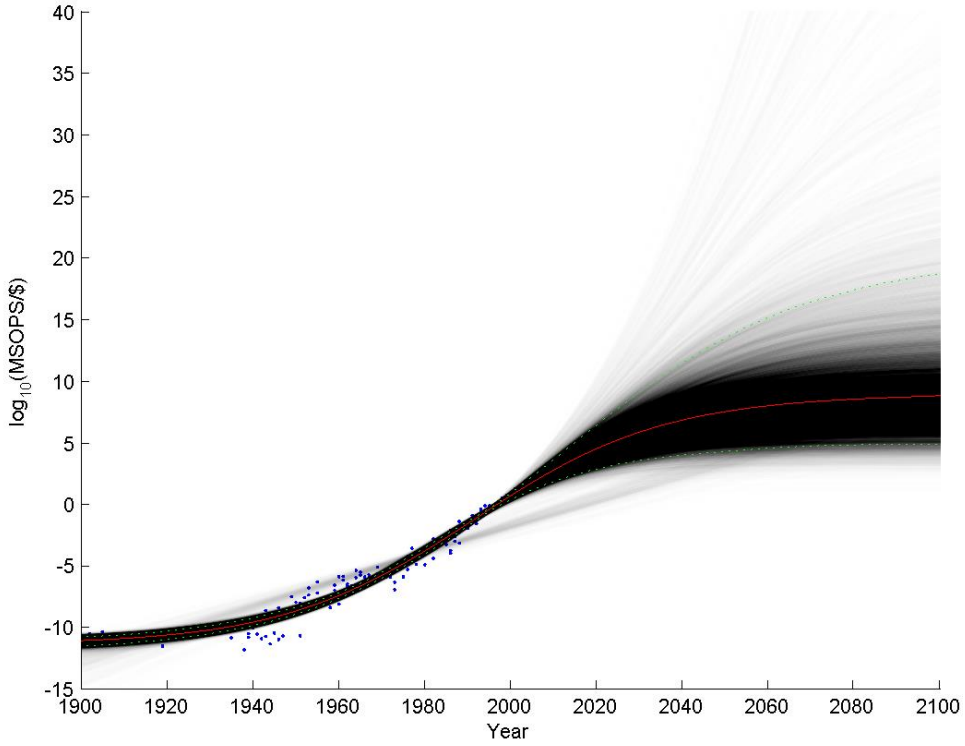

Figure 1: Distribution of scenarios of computer power generated from data in (Nordhaus 2001). The grey shade denotes density of scenarios. Red line is the median scenario, green lines the 5% and 95% percentile scenarios. The straight line

[10] These limits include the Margolus–Levitin theorem bounding the number of operations per second per joule of energy to below $h/4E \approx 10^{33}$ ops (Margolus & Levitin 1998) and the Bremermann limit $c^2/h \approx 10^{50}$ bits per second per kilogram (Bremermann 1962). (Lloyd 2000) gives more detail, and points out that a computer using individual atoms for bits and electromagnetic interactions can achieve up to 10^{40} operations per second per kilogram.

[11] The resampling consists of drawing *N* data points (with replacement) from the size *N* data set. This scenario generation method is based on jackknife sampling, a versatile statistical method where an estimator is repeatedly calculated from a set of randomly drawn members of the real data in order to estimate of the uncertainty in the estimator.

[12] (Nordhaus 2001)

fits (corresponding to fits with $c_3 \approx 0$), while visually salient, do not have an appreciable probability and hence do not affect the model results noticeably (if they did, they would cause a bias towards later arrival of WBE).

The median computing power in 2100 (when nearly all scenarios show significant levelling off) $6.4 \cdot 10^8$ MSOPS/$, with a 90% confidence interval between 77,000 and $4.9 \cdot 10^{18}$ MSOPS/$.

The intersection (if it exists) between the curve $M(t) \cdot B$ (the total computing power available for the project budget) and the requirements C is then used as an estimate of the time $T_{hardware}$ when the minimal amount of computing power is available[13].

$$M(T_{hardware}) \cdot B = C$$

Scanning

I model that at some point in time over the next 30 years research will start on scanning and scan interpretation[14].

$$T_{start} \sim 2014 + U(0,30)$$

Scanning is basically engineering and can likely be done on timescales comparable to projects such as HUGO, i.e. about a decade from the starting data. The time from the start of serious scanning experiments to a successful result is modelled as a normal distribution with mean 10 years and standard deviation 10 (negative results are negated).

$$T_{scanning} \sim T_{start} + \mathrm{abs}(N(10,10))$$

Again, this estimate should be refined by comparing to other research projects scaling up a measurement technology from a small scale to a large.

Neuroscience

Interpretation of scans into simulation needs to be done. This requires a combination of computer vision algorithms, neuroscience experimentation and validation methods. This is the hardest part to estimate, since it depends both on research funding, the fit between chosen research methods and the unknown reality, and to some extent luck and skill. I model that as a broader distribution (standard deviation 30 years) starting from the research start date.

$$T_{neuro} \sim T_{start} + \mathrm{abs}(N(10,30))$$

Note that this is based on an optimistic assumption that there is a solution to the scan interpretation problem: if for example brains are not computable, scanning cannot resolve relevant data, or other

[13] Note that this corresponds to one emulation running in real-time. Slower or faster emulations will require proportionally less or more computer power; a project willing to produce a slowed down version may hence achieve WBE earlier. This has not been modelled in this version of the model.

[14] Given current projects such as the US BRAIN project, the EU Human Brain Project, the NIH Human Connectome Project, and existing ventures in connectomics such as the ATLUM system of Lichtman and Hayworth at Harvard, KESM of 3Scan/Texas A&M, the EyeWire project of the Seung lab at MIT, some projects are clearly *already* underway. The model assumes a shift from the current exploratory small-scan paradigm at some point into a more industrial paradigm aiming for brain-scale scan.

key assumptions are false[15] then clearly no WBE will be forthcoming. However, this is a probability that cannot be estimated, and it does not change the arrival date of WBE if it *is* feasible.

Of more practical concern for the model is whether there is any way of estimating scientific progress on an open ended problem like scan interpretation. Most likely this is not possible, since time-to-success of comparable projects cannot be ascertained before learning much about what *kind* of project the interpretation problem turns out to be: the reference class will become apparent too late to be useful. This is different from the more concrete engineering project of scaling up the scan method. Further decomposition of the scan interpretation problem may help resolve some of the uncertainty.

WBE arrival

Finally, the earliest time WBE can be achieved given the particular scenario is when hardware, scanning and neuroscience have all arrived.

$$T_{WBE} = \max(T_{hardware}, T_{scanning}, T_{neuro})$$

[15] See (Sandberg 2014a) for an overview.

Simulation

The above model was implemented in Matlab and 100,000 scenarios were generated. Given the above assumptions the following distribution of WBE arrival dates emerge:

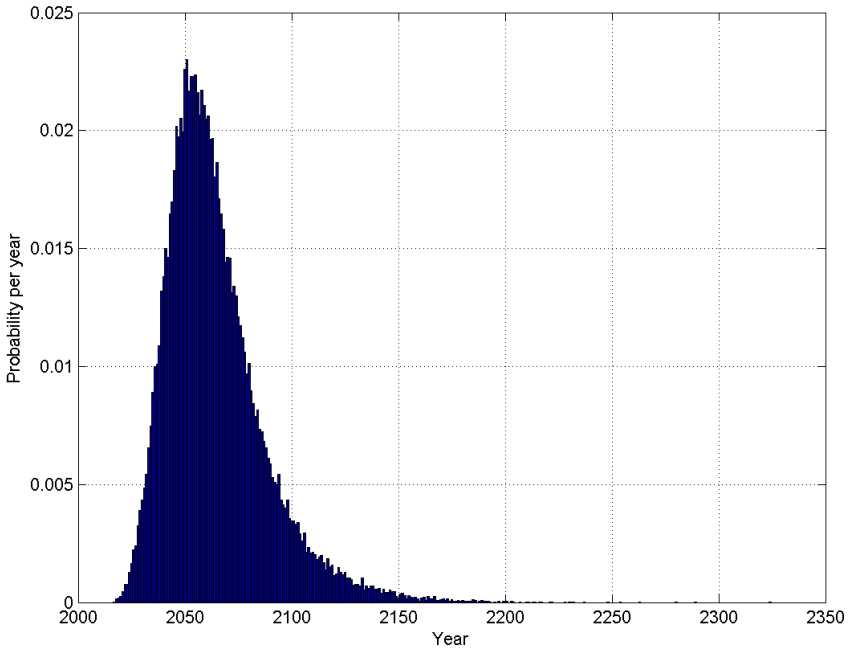

Figure 2: Estimated probability distribution for WBE arrival time.

Plotting the cumulative probability gives 50% chance for WBE (if it ever arrives) before 2059, with the 25% percentile in 2047 and the 75% percentile in 2074. WBE before 2030 looks very unlikely and only 10% likely before 2040.

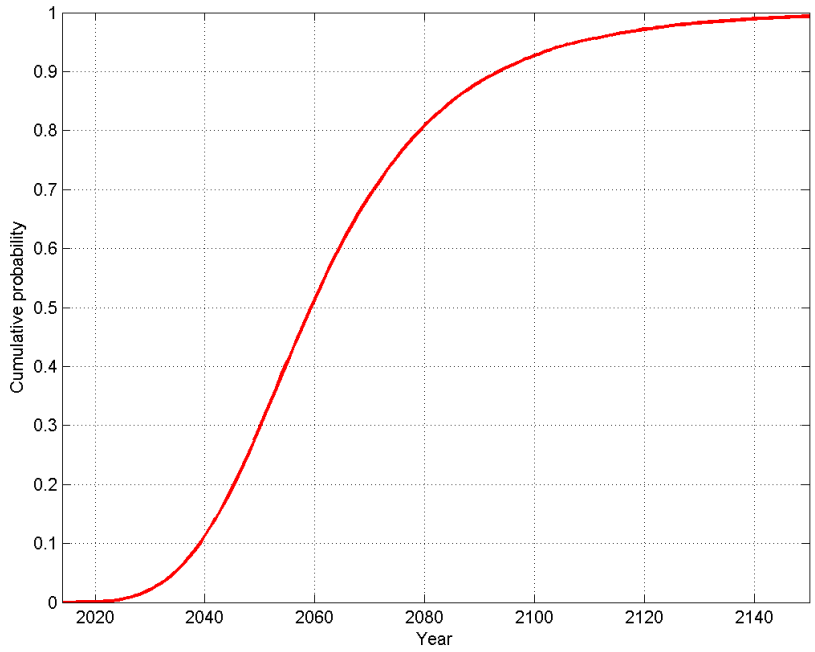

Figure 3: Cumulative probability distribution of WBE arrival time.

Requirements

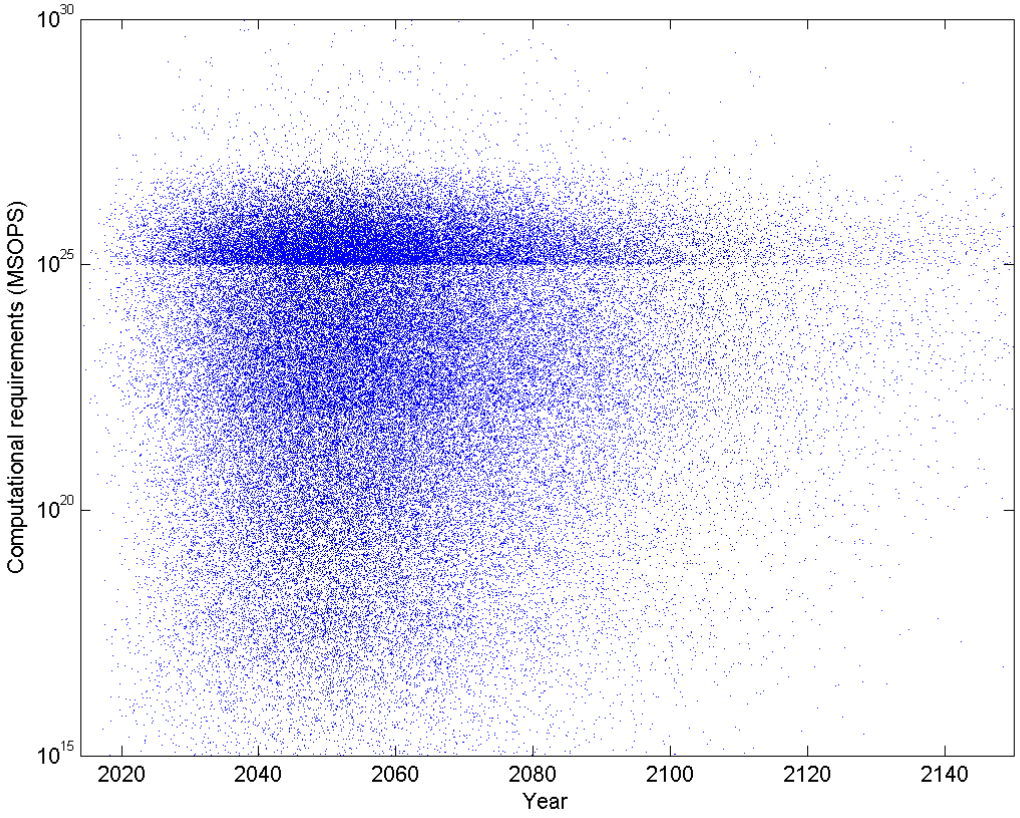

Figure 4: Scatterplot of the fundamental computational requirements versus the time of arrival for successful WBE scenarios. The horizontal lines are artefacts due to the discrete estimates of computational requirements.

If WBE requires more than 10^{27} MSOPS (roughly proteome level simulations), then it is unlikely to be feasible.

If WBE arrives early, it has significant spread in simulation requirements: early arrival can occur because emulation can be done at a crude resolution, but just as well because a project was fast, well-funded or lucky with Moore's law. Mid-range and late WBE has more probability mass for WBE being computationally hard.

Technology

The distribution of outcomes is split by the technology that turned out to be the final bottleneck. The colours denote whether hardware (blue), scanning (green) or neuroscience (red) arrives last[16].

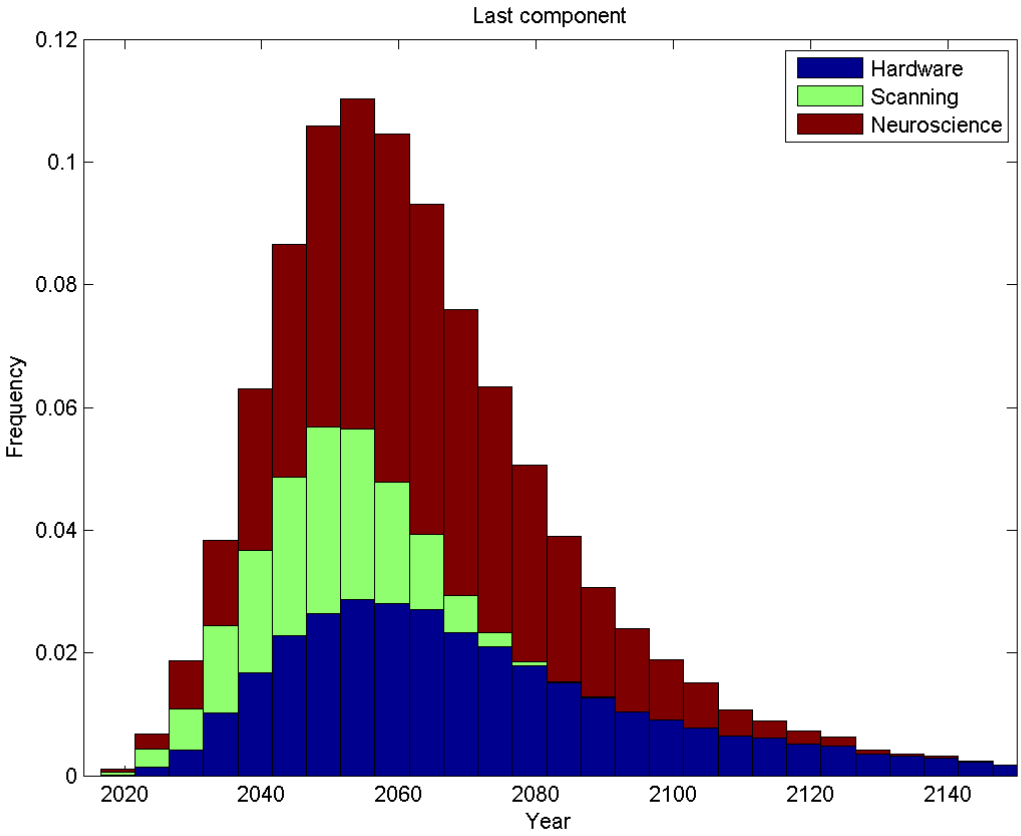

Figure 5: Distribution of scenarios where hardware (blue), scanning (green), or neuroscience (red) is the last key technology to arrive.

This distinction matters, because in hardware dominated scenarios there will be plenty of warning time before WBE becomes feasible during which smaller animal models show the feasibility yet scaling up directly to human level is not possible. In the case of neuroscience dominated scenarios breakthroughs can occur suddenly, perhaps while society at large regards the field as having made no progress over a long period and hence unlikely to be worth monitoring. The scan dominated scenarios may lead to situations where few brains are used as templates for many emulations.

It should be noted that neuroscience and scanning limited cases tend to occur somewhat earlier than the hardware limited cases, partially because of the research start within 30 year assumption and the given spread, partially because an extreme tail of hardware limited cases where the necessary computing requirements are just barely achieved by the fitted logistic function[17].

[16] 32% of scenarios are hardware-limited, 17% scanning-limited and 51% neuroscience limited. These probabilities are highly model-dependent.

[17] Given the assumptions, 51% of scenarios do not reach WBE due to hardware never becoming good enough to reach the proper brain scale. This should not be taken as a serious bound on probability of WBE, just the implication of current guesses in computational neuroscience and current trends in computing.

Overshoot

Overshoot measures how many how many simulations can be run for one million dollars when WBE arrives. A one million dollar human-level simulation is roughly in the ballpark where they become economically competitive with flesh-and-blood humans. If a large number of emulations are possible from the start, a dramatic economic shift is likely.

Similarly, if there is enough computer power available to run many emulations in parallel, it may also allow the running of very fast emulations (up to speed limits set by the parallelizability of brains on the available hardware). Again, a large overshoot implies the possibility of very fast emulations.

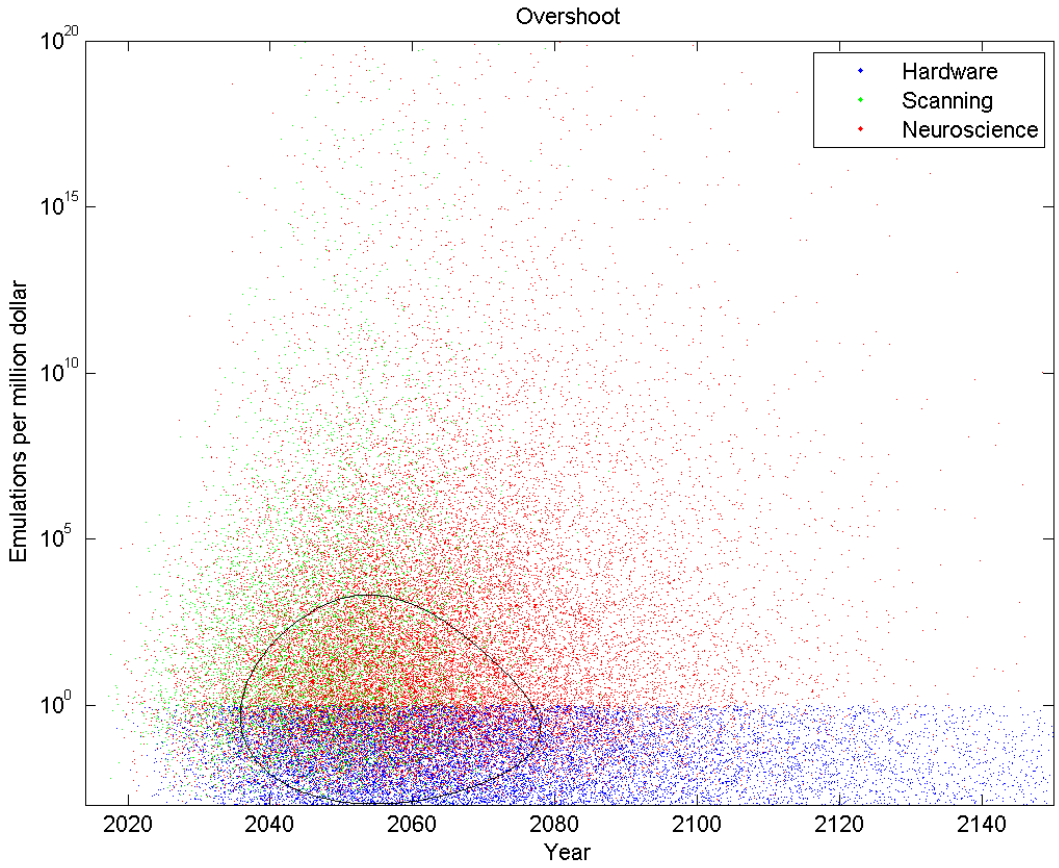

Figure 6: Scatterplot of scenarios showing the number of emulations that can be run for a million dollars. The contour curve represents 50% of maximum density. Color denotes which technology was last to arrive.

For scan and neuroscience-limited cases extreme overshoots are possible in the model. While the densest part of the diagram (the contour corresponds to 50% of the maximum probability density) have overshoots by less than a factor of 1,000 there is a wide spread; the 25% percentile (not shown) reaches a factor of several 100,000 and the 10% percentile a factor of 10^8.

The most extreme overshoots occur when the brain turns out to be unexpectedly simple and the breakthrough occurs after Moore's law has enabled a very large hardware overhang[18].

Meanwhile the hardware-limited case is by definition limited to an overshoot of a factor of 0.001 to 1; the smallest overshoots correspond to breakthroughs requiring a billion dollar project.

[18] This is somewhat similar to the scenario discussed in (Shulman & Sandberg 2010).

Discussion

This paper has demonstrated a simple model of WBE development and some of the conclusions that can be drawn from it. In particular, it shows that the probability of WBE being developed can go from negligible to geopolitically relevant over a span of a few decades. If success does not occur this century there may nevertheless remain a tail probability ≈8% for eventual success further on. Scanning-limited scenarios tend to be earlier than hardware- and neuroscience-limited scenarios. There is a noticeable potential for extreme overshoot even relatively early if there is a scan- or neuroscience-dominated breakthrough.

The model can be improved in many ways. Several possibilities have been described in the method section. For long-range futures there should be growing uncertainty in Moore's law, and the calibration should be updated with newer processor data. There should perhaps be a feedback between scanning and neuroscience, as good scanning methods are likely to accelerate neuroscience and vice versa. The assumptions about research start should be examined in the light of the progress of the newly started “big neuroscience” projects.

The predictions from this model should obviously be taken with a great deal of salt. In many ways it is merely a convolution of given prior distributions, in this case the author’s own guesses. But at least this is a model that allows exploration of the effects of different WBE assumptions in a transparent manner. As more information about the type and difficulty of the various projects needed arrive, it can be updated. Other approaches to WBE can be incorporated, ideally based on their roadmaps, to get a holistic picture of where the field is going.

Acknowledgments

I wish to thank Stuart Armstrong and Robin Hanson for encouraging the model and for forcing me to finish the write-up. Toby Ord originally suggested the intersection method for estimating earliest possible dates of WBE. Discussion with the original WBE roadmap workshop team, my FHI colleagues, MIRI, Peter Eckersley, and the networks of researchers actually making the technology real has helped shape my assumptions.

References

Armstrong, S. & Sotala, K. (2012) How We're Predicting AI—or Failing To. In *Beyond AI: Artificial Dreams*, edited by Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, and Radek Schuster, 52–75. Pilsen: University of West Bohemia. http://intelligence.org/files/PredictingAI.pdf

Bremermann, H.J. (1962) Optimization through evolution and recombination In: *Self-Organizing systems 1962*, edited M.C. Yovitts et al., Spartan Books, Washington, D.C. pp. 93–106.

Lloyd, S. (2000) Ultimate physical limits to computation. *Nature* 406 (6799): 1047–1054. http://arxiv.org/abs/quant-ph/9908043

Margolus, N. 7 Levitin, L. B. (1998) The maximum speed of dynamical evolution. *Physica D* 120: 188–195. http://arxiv.org/abs/quant-ph/9710043

Nordhaus, W. D. (2001) *The Progress of Computing*. Cowles Foundation Discussion Paper No. 1324. Available at SSRN: http://ssrn.com/abstract=285168

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

Sandberg, A. (2014a), 'Feasibility of whole brain emulation', in Vincent C. Müller (ed.), *Theory and Philosophy of Artificial Intelligence* (SAPERE; Berlin: Springer), 251-64. Sandberg, A. (2014) http://shanghailectures.org/sites/default/files/uploads/2013_Sandberg_Brain-Simulation_34.pdf

Sandberg, A. (2014b) Ethics of brain emulations. *Journal of Experimental & Theoretical Artificial Intelligence*, forthcoming 2014. http://www.aleph.se/papers/Ethics%20of%20brain%20emulations%20draft.pdf

Sandberg, A & Eckersley, P. (2014) Is Brain Emulation Dangerous? *Journal of Artificial General Intelligence*, forthcoming 2014

Shulman, C. & Sandberg, A. (2010) Implications of a Software-Limited Singularity. In *ECAP10: VIII European Conference on Computing and Philosophy*, edited by Klaus Mainzer. Munich: Dr. Hut. http://intelligence.org/files/SoftwareLimited.pdf