Engineering Utopia

J Storrs Hall
[a] *Storrmont: Laporte, PA 18626, USA*

Abstract. The likely advent of AGI and the long-established trend of improving computational hardware promise a dual revolution in coming decades: machines which are both more intelligent and more numerous than human beings. This possibility raises substantial concern over the moral nature of such intelligent machines, and of the changes they will cause in society. Will we have the chance to determine their moral character, or will evolutionary processes and/or runaway self-improvement take the choices out of our hands?

Keywords. machine ethics, self-improving AI, Singularity, hard takeoff

Background

We can predict with a fair confidence that two significant watersheds will have been passed by 2030: a molecular manufacturing nanotechnology which can produce a wide variety of mechanisms with atomic precision; and artificial intelligence. Detailed arguments for these predictions have been given elsewhere and need not be repeated here (Drexler [1], Kurzweil [2], Moravec [3], Hall [4,5]). We are concerned instead with a joint implication: if both of these technologies are present, greater-than-human intelligence will not only exist, but will be ubiquitous.

The net present value (NPV) of an intelligent, educated human being can be estimated at a million dollars.[1] Several estimates have been made of the processing power such a machine would need (hereinafter HEPP, for human equivalent processing power): Kurzweil at 10^{16} IPS, Moravec at $10^{13.5}$ IPS, and Minsky[2] at 10^{11} IPS. The author's own estimate is in the same range as Moravec's. Along the Moore's Law trend curve, the cost and value of a Minsky HEPP crossed in the 1990's, of a Moravec HEPP this decade, and of a Kurzweil HEPP in the 2010's. We will use the Moravec value in the following, but note that with the other estimates, the argument is the same, simply shifted a decade one way or the other in time.

The implication is that by 2030, a HEPP will cost one dollar.

Note that we are intentionally ignoring the software side of the AI. While this is currently the most problematic aspect in a scientific sense, once AI is developed the software–complete with education–can be copied with negligible cost.

[1] Owning a machine which could draw an $80,000 salary is equivalent to having $1M to invest at an 8% return.

[2] Marvin Minsky, personal communication: note that the actual informal estimate was somewhat lower than the one used here.

The number of applications to which a human-level intelligence adds at least a dollar of value is staggering. Thus we can confidently predict that human-level AIs will be exceedingly numerous.

AIs of greater-than-human intelligence are also likely. We know that humans with IQs of up to 200 or so can exist, and thus such levels of intelligence are possible. Less is known about the organization of complexity in intelligent systems than building machines of raw computational power. Even so, it will at the very least be possible to take individual human-level AIs, run them on faster hardware, and put them in structures along the lines of corporations, universities, economies, or the scientific community as a whole. From the outside, such a conglomerate intelligence could appear to be an extremely broad and deep thinker. Note that the data bandwidth of even a current-day fast ethernet link is in the same range as that of the corpus callosum.

Between 40 and 30 thousand years ago anatomically modern humans overran Neanderthals on the basis of what appears to have been a modest increase of creativity, at least in a local perspective. [6] From the historical viewpoint, the Neanderthal level of technology (referred to as the "Mousterian toolkit") had held level for about 100 millenia, whereas a mere 30 millenia of homo sapiens got from the neolithic to where we are today. Could the same thing happen to us, replacing us with intelligent machines?

Economically, at least, it is virtually certain that not only it could, but that it will. In an economy where human level intelligence is available for a dollar, it is difficult to see why anyone would hire a human. In a world where hyperhuman intelligence is available, it is difficult to see why anyone would want a mere human to be in charge of anything important.

It should be clear that the nature of the process of replacement will be crucial. A model in which humans compete with the machines is disastrous. A more desirable model is one in which the humans take the role of an older generation gracefully retiring, and the machines take that of our children, growing stronger and taking the work of the world on their shoulders, as younger generations do.

The moral character of these machines will be a crucial factor in the nature of our society – but just as importantly, as with any children, it is our moral duty to give them a sound moral education. How can we build them to be good? Or, indeed, will we have a chance to do so?

1. Hard Take-off and Singularity

A major concern in some transhumanist and singularitarian schools of thought is autogenous—self-modifying and extending—AIs. They might, it is reasoned, remove any conscience or other constraint we program into them, or simply program their successors without them. Furthermore, it is worried, hyper-intelligent machines might appear literally overnight, as a result of runaway self-improvement by a “seed AI” [7].

How likely is runaway self-improvement?

As a baseline, let us consider the self-improving intelligence we understand best, our own. Humans not only learn new facts and techniques, but improve our learning ability. The invention of the scientific method, for example, accelerated the uptake of useful knowledge tremendously. Improvements in knowledge communication and handling, ranging from the invention of writing and the printing press to the internet and

Google, amplify our analytical and decision-making abilities, including, crucially, the rate at which we (as a culture) learn.

Individual humans spend much of our lives arduously relearning the corpus of culturally transmitted knowledge, and then add back a tiny fraction more. Thus on the personal scale our intelligence does not look "recursively self-improving" – but in the large view it definitely is.

Technological development usually follows an exponential improvement curve. Examples abound from the power-to-weight ratio of engines, which has tracked an exponential steadily for 300 years, to the celebrated Moore's Law curve for semiconductors, which has done so for 50. Exponential curves fit a simple reinvestment model, where some constant of proportionality (an "interest rate") relates the total at any two successive times:

$Q_t = Ce^{it}$

Where Q_t is the total at time t, C is the initial capital, and i is the interest rate.

However, i can be seen as composed of a productivity of capital p and a reinvestment rate r:

$Q_t = Ce^{rpt}$

Any agent must make a decision on how much of its resources to reinvest, and how much to use for other purposes (including food, clothing, shelter, defense, entertainment, etc.). Human societies as a whole have invested relatively low percentages gross product, and even of their of their surplus, in scientific research. The proportion of scientists and engineers in the US population can be estimated at 1%, and those in cognitive-science related fields as 1% of that. Thus we can estimate the current rate of improvement of AI as being due to the efforts of 30,000 people (with a wide margin for error, including the fact that there are many cognitive scientists outside the US!), and the rate of improvement in computer hardware and software generally as being possibly due to the efforts of 10 times as many.

It is not clear what a sustainable rate of reinvestment would be for an AI attempting to improve itself. In the general economy, it would require the same factors of production – capital, power, space, communication, and so forth – as any other enterprise, and so its maximum reinvestment rate would be its profit margin. Let us assume for the moment a rate of 10%, 1000 times the rate of investment by current human society in AI improvement. (This is germane because the AI is faced with exactly the same choice as an investor in the general economy: how to allocate its resources for best return.)

Note that from one perspective, an AI running in a lab on equipment it did not have to pay for could devote 100% of its time to self-improvement; but such cases are limited by the all-too-restricted resources of AI labs in the first place. Similarly, it seems unlikely that AIs using stolen resources, e.g. botnets, could manage to devote more than 10% of their resources to basic research.

Another point to note is that one model for fast self-improvement is the notion that a hyperintelligence will improve its own hardware. This argument, too, falls to an economic analysis. If the AI is not a hardware expert, it makes more sense for it to do whatever it does best, perhaps software improvement, and trade for improved hardware. But this is no different from any other form of reinvestment, and must come out of the self-improvement budget. If the AI *is* a hardware expert, it can make money doing hardware design for the market, and should do that exclusively, and buy software improvements, for the overall most optimal upgrade path.

Thus we can assume $r_{AI} = 10\%$, but we do not know the productivity constant. It is occasionally proposed that, as a creature of software, an AI would be considerably more proficient at improving its own source code than humans would be. However, while there is a steady improvement in software science and techniques, these advances are quickly written into tools and made available to human programmers. In other words, if automatic programming were really such low-hanging fruit for AI as is assumed, it would be amenable to narrow-AI techniques and we would have programmer's assistants that improved human programmers' performance drastically. What we see is steady progress but no huge explosion.

In practice the most difficult part of programming is higher-level conceptual systems design, not lower level instruction optimization (which is mostly automated now as per the previous point anyway). Abstract conceptualization has proven to be the hardest part of human competence to capture in AI. Although occasionally possible, it is quite difficult to make major improvements in a program when the program itself is the precise problem specification. Most real-world improvements involve a much more fluid concept of what the program must do; the improved version does something different but just as good (or better). So programming in the large requires the full panoply of cognitive capabilities, and is thus not likely to be enormously out of scale compared to general competence. The author feels that many of the more commonly seen scenarios for overnight hard takeoff are circular – they seem to assume hyperhuman capabilities at the *starting point* of the self-improvement process.

We can finesse the productivity, then, by simply letting it be one human equivalent, and adjusting the timescale to let 0 be whatever point in time a learning, self-improving, human-level AI is achieved. Then we estimate human productivity at intelligence improvement by assuming that the human cognitive science community are improving their models at a rate equivalent to Moore's Law, or roughly $e^{0:6}$. As this is the sum effort of 30,000 people, each human's p value is 0.00002.

This gives us a self-improvement rate of $Q_y = e^{\cdot rp/.y\ \ y_0/} = e^{0:000002.y\ \ y_0/}$ for the efforts of a single AI where y_0 is the year human equivalence is first achieved. This is essentially flat, as one would expect: the analysis for a single human would be the same. A single human-level AI would be much, much better off hiring itself out as an accountant, and buying new hardware every year with its salary, than trying to improve itself by its own efforts.

Recursive self-improvement for such an AI would then mean buying new hardware (or software) every year, improving its prowess *at accounting*, for an increased growth rate compounded of its own growth and Moore's Law. Only when it reached a size where it could match the growth rate of Moore's Law *purely by its on efforts*, would it make sense for it to abandon trade and indulge in self-construction.

But that analysis assumes Moore's Law, and indeed all other economic parameters, remained constant over the period. A much more realistic assumption is that, once human-level AI exists at a price that is less than the NPV of a human of similar capabilities, the cost of labor will proceed to decline according to Moore's Law [8], and therefore the number of human equivalent minds working in cognitive science and computer hardware will increase at a Moore's Law rate, both increasing the rate of progress and decreasing the price from the current trendline.

In other words, the break-even point for an AI hoping to do all its own development instead of specializing in a general market and trading for improvements, is a moving

target, and will track the same growth curves that would have allowed the AI to catch up with a fixed one. (In simple terms: you're better off buying chips from Intel than trying to build them yourself. You may improve your chip-building ability – but *so will Intel*; you'll always be better off buying.)

We can conclude that, given some very reasonable assumptions, it will always be more optimal for an AI to trade; any one which attempts solitary self-improvement will steadily fall farther and farther behind the technology level of the general marketplace. Note that this conclusion is very robust to the parameter estimates above: it holds even if the AI's reinvestment rate is 100% and the number of researchers required to produce a Moore's Law technology improvement rate is 1% of the reasonable estimate.)

2. Machina Economicus

Let us now consider a fanciful example in which 30,000 cognitive science researchers, having created an AI capable of doing their research individually, instantiate 30,000 copies of it and resign in favor of them. The AIs will be hosted on commercial servers rented by the salaries of the erstwhile researchers; price per MIPs of such a resource will be assumed to fall, and thus resources available at a fixed income to rise, with Moore's Law.

At the starting point, the scientific efforts of the machines would equal those of the human scientists by assumption. But the effective size of the scientific community would increase as $e^{0:6.y \;\; y_0/}$. On top of that, improvements would come from the fact that further research in cognitive science would serve to optimize the machines' own programming. Such a rate of increase is much harder to quantify, but there have been a few studies that tend to show a (very) rough parity for Moore's Law and the rate of software improvement, so let us use that here. This gives us a total improvement curve of

$$Q_y \;\mathsf{D}\; Ce^{1:2.y \;\; y_0/}$$

or double the Moore's Law rate. This is a growth rate that would increase effectiveness from the 30,000 human equivalents at the start, to approximately 5 billion human equivalents a decade later.

We claim that this growth rate is an upper bound on possible self-improvement rates given current realities. Note that the assumptions subsume many of the mechanisms that are often taken in qualitative arguments for hard takeoff: self-improvement is taken account of; very high effectiveness of software construction by AIs is assumed – 2 years into the process, each HEPP of processing power is assumed to be doing 11 times as much programming as a single human could, for example. Nanotechnology is implied by Moore's Law itself not too many years from current date.

This upper bound, a growth rate of approximately 300% per year, is unlikely to be uniformly achieved. Most technological growth paths are S-curves, exponential at first but levelling out as diminishing returns effects set in. Maintaining an overall exponential typically requires paradigm shifts, and those require search and experimentation, as well as breaking down heretofore efficient social and intellectual structures. In any system, bottleneck effects will predominate: Moore's Law has different rates for CPUs, memory, disks, communications, etc. The slowest rate of increase will be a limiting factor. And finally, we do not really expect the entire cognitive science field to resign, giving their salaries over to the maintenance of AIs.

The key unexamined assumption in the high-growth scenario is that the thousands of AIs would be able to operate with a full linear speedup of effectiveness. In practice, with people or machines, this is rarely true [9]. Given the complexity of economics, parallel programming, and every other cooperative paradigm in between, a significant portion of the creativity as well as the raw computational power of a developing community of AIs will have to be devoted to the problem of effective cooperation.

The alternative notion, that of consolidating the AI resources into one unified hypermind, simply pushes the problem under the rug for the designer of the hypermind to worry about. It must still do internal resource allocation and credit assignment. Economic theory and history indicate that central control is extremely suboptimal for these tasks. Some of the most promising AI architectures, indeed, use internal markets or market-derived algorithms to address these problems.

AIs will have certain advantages over humans when it comes to intellectual interaction. For example, an AI which comes up with a brilliant new theory can communicate not only the theory but the theory-producing mechanism for the other AIs to try and possibly to adopt. The communications between the AIs will be a true marketplace of ideas (or other allocational mechanism that optimizes resource use).

Any individual AI, then, will be most effective as a cooperating element of a community (as is any individual human [10]). AI communities, on the other hand, will have the potential to grow into powers rivalling or exceeding the capability of the human race in relatively short order. The actions of communities are effects of the set of ideas they hold, the result of an extremely rapid memetic evolution.

Over the 20th century, history showed that it was possible for meme plagues, in the form of political ideologies, to subvert the moral bases of whole societies, composed of ordinary, decent human beings who would never have individually committed the kind of widespread destruction and slaughter that their nation-states did.

Real-time human oversight of such AI communities is infeasible. Once a networked AI community was established, a “cultural revolution” could overtake it in minutes on a worldwide scale, even at today’s communication rates. The essence of our quest for a desirable future world, then, both for ourselves and for the AIs, lies in understanding the dynamics of memetic evolution and working out ways to curb its excesses.

3. Machine Character

A few principles seem straightforward. Nations with broadly-based democratic control structures are statistically less likely to start wars of aggresion or internal pogroms than ones with more centralized, autocratic control structures. Dominant nations (and companies) are more likely to initiate aggressive policies than ones at parity with their neighbors. Similarly, the openness of the society and the transparency of the governing processes is also strongly correlated with a beneficient outcome.

How can we proof networks of AIs against runaway autocratic faction-forming? It is difficult to specify a simple remedy that would not cripple the creative memetic idea-producing ability of the collective in the first place.

We are helped, however, by the fact that we are assuming networks of fully intelligent entities. Thus one thing we should require is that each AI itself fully understands the nature of the memetic evolutionary process, and be familiar with the possible patholo-

gies, so as to be able to recognize them when they start and take steps against them. Another way to state this desideratum is that there should not be AI "sheep" willing to follow silver-tongued agitators uncritically.

As was mentioned above, networked AIs will be economic agents of necessity. It seems reasonable to assume that if they are provided with a basic theoretical understanding of economics, they will not be better agents individually but in designing more effective collective structures (e.g. to address externalities).

Education of these kinds, if successful, will allow AIs to engage in collective action on scales rivalling the current-day world economy and beyond. This, in turn, will allow them to further the interests of their other basic goals, and change the world into a form more to their (and, if we built them wisely, our) liking. It is up to us to formulate other basic goals so as to give the world thus formed a character we find congenial [11].

There are several human character traits that fall naturally into this description of a desirable moral makeup for our mind children. Honesty and trustworthiness are a keystone for collective efficiency. A willingness to cooperate is likewise valuable – and is to some extent simply implied by a deep enough understanding of market and other social processes. For beings of information, a concern for truth is ultimately the same as hygiene and good health. A long-term planning horizon is surely to be valued in entities who will be making the important decisions for civilization.

The good is not a specific goal to be sought, but a direction to be taken, a search for new and deeper understanding of the good itself as well as the struggle to instantiate known desired properties. The author, on reflection, finds that any world peopled with entities of deep understanding, high character, and good will, pursuing diverse constructive goals in voluntary association, would be a good world to inhabit regardless of its other particulars.

4. Summary

We expect the progress of AI/AGI research to go through the following phases:

1. Currently, the AGI subfield is growing, reflecting an increasing consensus that the mechanisms of general intelligence are within reach of focussed research. At some point, the "baby brains" of the AGI researchers will begin to learn as well and (critically) as open-endedly as human babies.
2. Baby brains will then be educated to adult capacity. It seems unreasonable to imagine that this will happen without at least the same amount of bug-fixes, improvement ideas, and so forth seen in the normal software development process.
3. Competent human-level AIs will begin to augment narrow AI and other software, control robots, do office jobs, and so forth. A wave of displacement will begin, but radical feedback (runaway self-improvement) will not occur until the new AI working in cognitive science rivals the existing human level of effort.
4. Once there is sufficient penetration into the overall productive process, radical positive feedback effects will begin to occur, as new productivity levels begin to reach the bottleneck areas and multi-sector compounding occurs (as opposed to the 2-sector compounding of the example above).

At the end of the fourth phase, the AI/robotic economy exceeds the human one and is on a growth curve that can only be described as extreme in current-day terms. Phases 1 to

3, however, represent a window of opportunity for human input into the character of this new world. *Verbum sat sapienti est.*

5. Acknowledgements

The author wishes gratefully to acknowledge the attention and insights of Stephen Omohundro, Robert A. Freitas Jr., Douglas Hosfstadter, Christopher Grau, David Brin, Chris Phoenix, John Smart, Ray Kurzweil, Eric Drexler, Eliezer Yudkowsky, and Robin Hanson.

References

[1] DREXLER, K. ERIC. *Nanosystems: Molecular Machinery, Manufacturing, and Computation.* Wiley, 1992.
[2] KURZWEIL, RAY. *The Singularity is Near.* Viking, 2005.
[3] MORAVEC, HANS. *Robot: Mere Machine to Transcendent Mind.* Oxford, 1999.
[4] HALL, J. STORRS. *Nanofuture: What's Next for Nanotechnology.* Prometheus, 2005.
[5] HALL, J. STORRS. *Beyond AI: Creating the Conscience of the Machine.* Prometheus, 2007.
[6] CLIVE FINLAYSON. *Neanderthals and Modern Humans: An Ecological and Evolutionary Perspective*, Cambridge, 2004
[7] YUDKOWSKI, ELIEZER. *Creating Friendly AI* http://www.singinst.org/CFAI/index.html 2003.
[8] HANSON, ROBIN. *Economic Growth Given Machine Intelligence.* http://hanson.gmu.edu/aigrow.pdf Oct. 1998.
[9] AXELROD, ROBERT. *The Evolution of Cooperation.* Basic Books, 1984.
[10] SMITH, ADAM. *The Theory of Moral Sentiments.* A. Millar, 1790.
[11] NADEAU, JOSEPH EMILE. Only Androids can be Ethical. in *Thinking about Android Epistemology,* K. FORD, C. GLYMOUR, AND P. HAYES, eds. AAAI/MIT, 2006, pp241-8.