3. From AI to Machine Superintelligence

It seems unlikely that humans are near the ceiling of possible intelligences, rather than simply being the first such intelligence that happened to evolve. Computers far outperform humans in many narrow niches (e.g. arithmetic, chess, memory size), and there is reason to believe that similar large improvements over human performance are possible for general reasoning, technology design, and other tasks of interest. As occasional AI critic Jack Schwartz (1987) wrote:

> If artificial intelligences can be created at all, there is little reason to believe that initial successes could not lead swiftly to the construction of artificial superintelligences able to explore significant mathematical, scientific, or engineering alternatives at a rate far exceeding human ability, or to generate plans and take action on them with equally overwhelming speed. Since man's near-monopoly of all higher forms of intelligence has been one of the most basic facts of human existence throughout the past history of this planet, such developments would clearly create a new economics, a new sociology, and a new history.

Why might AI "lead swiftly" to machine superintelligence? Below we consider some reasons.

3.1. AI Advantages

Below we list a few AI advantages that may allow AIs to become not only vastly more intelligent than any human, but also more intelligent than all of biological humanity (Sotala 2012; Legg 2008). Many of these are unique to *machine* intelligence, and that is why we focus on intelligence explosion from AI rather than from biological cognitive enhancement (Sandberg 2011).

Increased computational resources. The human brain uses 85–100 billion neurons. This limit is imposed by evolution-produced constraints on brain volume and metabolism. In contrast, a machine intelligence could use scalable computational resources (imagine a "brain" the size of a warehouse). While algorithms would need to be changed in order to be usefully scaled up, one can perhaps get a rough feel for the potential impact here by noting that humans have about 3.5 times the brain size of chimps (Schoenemann 1997), and that brain size and IQ correlate positively in humans, with a correlation coefficient of about 0.35 (McDaniel 2005). One study suggested a similar correlation between brain size and cognitive ability in rats and mice (Anderson 1993).[15]

15. Note that given the definition of intelligence we are using, greater computational resources would not give a machine more "intelligence" but instead more "optimization power."

Communication speed. Axons carry spike signals at 75 meters per second or less (Kandel, Schwartz, and Jessell 2000). That speed is a fixed consequence of our physiology. In contrast, software minds could be ported to faster hardware, and could therefore process information more rapidly. (Of course, this also depends on the efficiency of the algorithms in use; faster hardware compensates for less efficient software.)

Increased serial depth. Due to neurons' slow firing speed, the human brain relies on massive parallelization and is incapable of rapidly performing any computation that requires more than about 100 sequential operations (Feldman and Ballard 1982). Perhaps there are cognitive tasks that could be performed more efficiently and precisely if the brain's ability to support parallelizable pattern-matching algorithms were supplemented by support for longer sequential processes. In fact, there are many known algorithms for which the best parallel version uses far more computational resources than the best serial algorithm, due to the overhead of parallelization.[16]

Duplicability. Our research colleague Steve Rayhawk likes to describe AI as "instant intelligence; just add hardware!" What Rayhawk means is that, while it will require extensive research to design the first AI, creating additional AIs is just a matter of copying software. The population of digital minds can thus expand to fill the available hardware base, perhaps rapidly surpassing the population of biological minds.

Duplicability also allows the AI population to rapidly become dominated by newly built AIs, with new skills. Since an AI's skills are stored digitally, its exact current state can be copied,[17] including memories and acquired skills—similar to how a "system state" can be copied by hardware emulation programs or system backup programs. A human who undergoes education increases only his or her own performance, but an AI that becomes 10% better at earning money (per dollar of rentable hardware) than other AIs can be used to replace the others across the hardware base—making each copy 10% more efficient.[18]

Editability. Digitality opens up more parameters for controlled variation than is possible with humans. We can put humans through job-training programs, but we can't perform precise, replicable neurosurgeries on them. Digital workers would be more editable than human workers are. Consider first the possibilities from whole brain emulation. We know that transcranial magnetic stimulation (TMS) applied to one part of

16. For example see Omohundro (1987).

17. If the first self-improving AIs at least partially require quantum computing, the system states of these AIs might not be directly copyable due to the no-cloning theorem (Wootters and Zurek 1982).

18. Something similar is already done with technology-enabled business processes. When the pharmacy chain CVS improves its prescription-ordering system, it can copy these improvements to more than 4,000 of its stores, for immediate productivity gains (McAfee and Brynjolfsson 2008).

the prefrontal cortex can improve working memory (Fregni et al. 2005). Since TMS works by temporarily decreasing or increasing the excitability of populations of neurons, it seems plausible that decreasing or increasing the "excitability" parameter of certain populations of (virtual) neurons in a digital mind would improve performance. We could also experimentally modify dozens of other whole brain emulation parameters, such as simulated glucose levels, undifferentiated (virtual) stem cells grafted onto particular brain modules such as the motor cortex, and rapid connections across different parts of the brain.[19] Secondly, a modular, transparent AI could be even more directly editable than a whole brain emulation—possibly via its source code. (Of course, such possibilities raise ethical concerns.)

Goal coordination. Let us call a set of AI copies or near-copies a "copy clan." Given shared goals, a copy clan would not face certain goal coordination problems that limit human effectiveness (J. W. Friedman 1994). A human cannot use a hundredfold salary increase to purchase a hundredfold increase in productive hours per day. But a copy clan, if its tasks are parallelizable, could do just that. Any gains made by such a copy clan, or by a human or human organization controlling that clan, could potentially be invested in further AI development, allowing initial advantages to compound.

Improved rationality. Some economists model humans as *Homo economicus*: self-interested rational agents who do what they believe will maximize the fulfillment of their goals (M. Friedman 1953). On the basis of behavioral studies, though, Schneider (2010) points out that we are more akin to Homer Simpson: we are irrational beings that lack consistent, stable goals (Schneider 2010; Cartwright 2011). But imagine if you *were* an instance of *Homo economicus*. You could stay on a diet, spend the optimal amount of time learning which activities will achieve your goals, and then follow through on an optimal plan, no matter how tedious it was to execute. Machine intelligences of many types could be written to be vastly more rational than humans, and thereby accrue the benefits of rational thought and action. The rational agent model (using Bayesian probability theory and expected utility theory) is a mature paradigm in current AI design (Hutter 2005; Russell and Norvig 2010, ch. 2).

These AI advantages suggest that AIs will be *capable* of far surpassing the cognitive abilities and optimization power of humanity as a whole, but will they be *motivated* to do so? Though it is difficult to predict the specific motivations of advanced AIs, we can make some predictions about convergent instrumental goals—instrumental goals useful for the satisfaction of almost any final goals.

19. Many suspect that the slowness of cross-brain connections has been a major factor limiting the usefulness of large brains (Fox 2011).

3.2. Instrumentally Convergent Goals

Omohundro (2007, 2008, 2012) and Bostrom (forthcoming) argue that there are several instrumental goals that will be pursued by almost any advanced intelligence because those goals are useful intermediaries to the achievement of almost any set of final goals. For example:

1. An AI will want to preserve itself because if it is destroyed it won't be able to act in the future to maximize the satisfaction of its present final goals.
2. An AI will want to preserve the content of its current final goals because if the content of its final goals is changed it will be less likely to act in the future to maximize the satisfaction of its present final goals.[20]
3. An AI will want to improve its own rationality and intelligence because this will improve its decision-making, and thereby increase its capacity to achieve its goals.
4. An AI will want to acquire as many resources as possible, so that these resources can be transformed and put to work for the satisfaction of the AI's final and instrumental goals.

Later we shall see why these convergent instrumental goals suggest that the default outcome from advanced AI is human extinction. For now, let us examine the mechanics of AI self-improvement.

3.3. Intelligence Explosion

The convergent instrumental goal for self-improvement has a special consequence. Once human programmers build an AI with a better-than-human *capacity* for AI design, the instrumental goal for self-improvement may motivate a positive feedback loop of self-enhancement.[21] Now when the machine intelligence improves itself, it improves the intelligence that does the improving. Thus, if mere human efforts suffice to produce machine intelligence this century, a large population of greater-than-human machine intelligences may be able to create a rapid cascade of self-improvement cycles, enabling

20. Bostrom (2012) lists a few special cases in which an AI may wish to modify the content of its final goals.

21. When the AI can perform 10% of the AI design tasks and do them at superhuman speed, the remaining 90% of AI design tasks act as bottlenecks. However, if improvements allow the AI to perform 99% of AI design tasks rather than 98%, this change produces a much larger impact than when improvements allowed the AI to perform 51% of AI design tasks rather than 50% (Hanson 1998). And when the AI can perform 100% of AI design tasks rather than 99% of them, this removes altogether the bottleneck of tasks done at slow human speeds.

a rapid transition to machine superintelligence. Chalmers (2010) discusses this process in some detail, so here we make only a few additional points.

The term "self," in phrases like "recursive self-improvement" or "when the machine intelligence improves itself," is something of a misnomer. The machine intelligence could conceivably edit its own code while it is running (Schmidhuber 2007; Schaul and Schmidhuber 2010), but it could also create new intelligences that run independently. Alternatively, several AIs (perhaps including WBEs) could work together to design the next generation of AIs. Intelligence explosion could come about through "self"-improvement or through other-AI improvement.

Once sustainable machine self-improvement begins, AI development need not proceed at the normal pace of human technological innovation. There is, however, significant debate over how fast or local this "takeoff" would be (Hanson and Yudkowsky 2008; Loosemore and Goertzel 2011; Bostrom, forthcoming), and also about whether intelligence explosion would result in a stable equilibrium of multiple machine superintelligences or instead a machine "singleton" (Bostrom 2006). We will not discuss these complex issues here.

4. Consequences of Machine Superintelligence

If machines greatly surpass human levels of intelligence—that is, surpass humanity's capacity for efficient cross-domain optimization—we may find ourselves in a position analogous to that of the apes who watched as humans invented fire, farming, writing, science, guns and planes and then took over the planet. (One salient difference would be that no single ape witnessed the entire saga, while we might witness a shift to machine dominance within a single human lifetime.) Such machines would be superior to us in manufacturing, harvesting resources, scientific discovery, social aptitude, and strategic action, among other capacities. We would not be in a position to negotiate with them, just as neither chimpanzees nor dolphins are in a position to negotiate with humans.

Moreover, intelligence can be applied in the pursuit of any goal. As Bostrom (2012) argues, making AIs more intelligent will not make them want to change their goal systems—indeed, AIs will be motivated to *preserve* their initial goals. Making AIs more intelligent will only make them more capable of achieving their original final goals, whatever those are.[22]

This brings us to the central feature of AI risk: Unless an AI is specifically programmed to preserve what humans value, it may destroy those valued structures (in-

22. This may be less true for early-generation WBEs, but Omohundro (2007) argues that AIs will converge upon being optimizing agents, which exhibit a strict division between goals and cognitive ability.

cluding humans) *incidentally*. As Yudkowsky (2008a) puts it, "the AI does not love you, nor does it hate you, but you are made of atoms it can use for something else."

4.1. Achieving a Controlled Intelligence Explosion

How, then, can we give AIs desirable goals before they self-improve beyond our ability to control them or negotiate with them?[23] WBEs and other brain-inspired AIs running on human-derived "spaghetti code" may not have a clear "slot" in which to specify desirable goals (Marcus 2008). The same may also be true of other "opaque" AI designs, such as those produced by evolutionary algorithms—or even of more transparent AI designs. Even if an AI had a transparent design with a clearly definable utility function,[24] would we know how to give it desirable goals? Unfortunately, specifying what humans value may be extraordinarily difficult, given the complexity and fragility of human preferences (Yudkowsky 2011; Muehlhauser and Helm 2012), and allowing an AI to *learn* desirable goals from reward and punishment may be no easier (Yudkowsky 2008a). If this is correct, then the creation of self-improving AI may be detrimental *by default* unless we first solve the problem of how to build an AI with a stable, desirable utility function—a "Friendly AI" (Yudkowsky 2001).[25]

But suppose it is possible to build a Friendly AI (FAI) capable of radical self-improvement. Normal projections of economic growth allow for great discoveries relevant to human welfare to be made eventually—but a Friendly AI could make those discoveries much sooner. A benevolent machine superintelligence could, as Bostrom (2003) writes, "create opportunities for us to vastly increase our own intellectual and emotional capabilities, and it could assist us in creating a highly appealing experiential world in which we could live lives devoted [to] joyful game-playing, relating to each other, experiencing, personal growth, and to living closer to our ideals."

---

23. Hanson (2012) reframes the problem, saying that "we should expect that a simple continuation of historical trends will eventually end up [producing] an 'intelligence explosion' scenario. So there is little need to consider [Chalmers'] more specific arguments for such a scenario. And the inter-generational conflicts that concern Chalmers in this scenario are generic conflicts that arise in a wide range of past, present, and future scenarios. Yes, these are conflicts worth pondering, but Chalmers offers no reasons why they are interestingly different in a 'singularity' context." We briefly offer just one reason why the "inter-generational conflicts" arising from a transition of power from humans to superintelligent machines are interestingly different from previous the inter-generational conflicts: as Bostrom (2002) notes, the singularity may cause the extinction not just of people groups but of the entire human species. For a further reply to Hanson, see Chalmers (2012).

24. A utility function assigns numerical utilities to outcomes such that outcomes with higher utilities are always preferred to outcomes with lower utilities (Mehta 1998).

25. It may also be an option to constrain the first self-improving AIs just long enough to develop a Friendly AI before they cause much damage.