
# Intelligence Explosion: Evidence and Import

Luke Muehlhauser, Anna Salamon
*Machine Intelligence Research Institute*

## Abstract

In this chapter we review the evidence for and against three claims: that (1) there is a substantial chance we will create human-level AI before 2100, that (2) if human-level AI is created, there is a good chance vastly superhuman AI will follow via an "intelligence explosion," and that (3) an uncontrolled intelligence explosion could destroy everything we value, but a controlled intelligence explosion would benefit humanity enormously if we can achieve it. We conclude with recommendations for increasing the odds of a controlled intelligence explosion relative to an uncontrolled intelligence explosion.

Muehlhauser, Luke, and Anna Salamon. 2012. "Intelligence Explosion: Evidence and Import." In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, edited by Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. Berlin: Springer.

This version contains changes.

> The best answer to the question, "Will computers ever be as smart as humans?" is probably "Yes, but only briefly."
>
> —Vernor Vinge

## 1. Introduction

Humans may create human-level[1] artificial intelligence (AI) this century. Shortly thereafter, we may see an "intelligence explosion" or "technological singularity"—a chain of events by which human-level AI leads, fairly rapidly, to intelligent systems whose capabilities far surpass those of biological humanity as a whole.

How likely is this, and what will the consequences be? Others have discussed these questions previously (Turing 1950, 1951; Good 1959, 1965, 1970, 1982; von Neumann 1966; Minsky 1984; Solomonoff 1985; Vinge 1993; Yudkowsky 2008a; Nilsson 2009, chap. 35; Chalmers 2010; Hutter 2012); our aim is to provide a brief review suitable both for newcomers to the topic and for those with some familiarity with the topic but expertise in only *some* of the relevant fields.

For a more comprehensive review of the arguments, we refer our readers to Chalmers (2010, 2012) and Bostrom (forthcoming). In this short chapter we will quickly survey some considerations for and against three claims:

1. There is a substantial chance we will create human-level AI before 2100;
2. If human-level AI is created, there is a good chance vastly superhuman AI will follow via an intelligence explosion;
3. An uncontrolled intelligence explosion could destroy everything we value, but a *controlled* intelligence explosion would benefit humanity enormously if we can achieve it.

Because the term "singularity" is popularly associated with several claims and approaches we will not defend (Sandberg 2010), we will first explain what we are *not* claiming.

First, we will not tell detailed stories about the future. Each step of a story may be probable, but if there are many such steps, the whole story itself becomes improbable (Nordmann 2007; Tversky and Kahneman 1983). We will not assume the continuation of Moore's law, nor that hardware trajectories determine software progress, nor that faster computer speeds necessarily imply faster "thought" (Proudfoot and Copeland 2012), nor that technological trends will be exponential (Kurzweil 2005) rather than "S-curved" or otherwise (Modis 2012), nor indeed that AI progress will accelerate rather

1. We will define "human-level AI" more precisely later in the chapter.

than decelerate (Plebe and Perconti 2012). Instead, we will examine convergent outcomes that—like the evolution of eyes or the emergence of markets—can come about through any of several different paths and can gather momentum once they begin. Humans tend to underestimate the likelihood of outcomes that can come about through many different paths (Tversky and Kahneman 1974), and we believe an intelligence explosion is one such outcome.

Second, we will not assume that human-level intelligence can be realized by a classical Von Neumann computing architecture, nor that intelligent machines will have internal mental properties such as consciousness or human-like "intentionality," nor that early AIs will be geographically local or easily "disembodied." These properties are not required to build AI, so objections to these claims (Lucas 1961; Dreyfus 1972; Searle 1980; Block 1981; Penrose 1994; van Gelder and Port 1995) are not objections to AI (Chalmers 1996, chap. 9; Nilsson 2009, chap. 24; McCorduck 2004, chap. 8 and 9; Legg 2008; Heylighen 2012) or to the possibility of intelligence explosion (Chalmers 2012).[2] For example: a machine need not be *conscious* to intelligently reshape the world according to its preferences, as demonstrated by goal-directed "narrow AI" programs such as the leading chess-playing programs.

We must also be clear on what we mean by "intelligence" and by "AI." Concerning "intelligence," Legg and Hutter (2007) found that definitions of intelligence used throughout the cognitive sciences converge toward the idea that "Intelligence measures an agent's ability to achieve goals in a wide range of environments." We might call this the "optimization power" concept of intelligence, for it measures an agent's power to optimize the world according to its preferences across many domains. But consider two agents which have equal ability to optimize the world according to their preferences, one of which requires much more computational time and resources to do so. They

2. Chalmers (2010) suggested that AI will lead to intelligence explosion if an AI is produced by an "extendible method," where an extendible method is "a method that can easily be improved, yielding more intelligent systems." McDermott (2012a, 2012b) replies that if P≠NP (see Goldreich [2010] for an explanation) then there is no extendible method. But McDermott's notion of an extendible method is not the one essential to the possibility of intelligence explosion. McDermott's formalization of an "extendible method" requires that the program generated by each step of improvement under the method be able to solve in polynomial time all problems in a particular class—the class of solvable problems of a given (polynomially step-dependent) size in an NP-complete class of problems. But this is not required for an intelligence explosion in Chalmers' sense (and in our sense). What intelligence explosion (in our sense) would require is merely that a program self-improve to *vastly outperform humans*, and we argue for the plausibility of this in section 3 of our chapter. Thus while we agree with McDermott that it is probably true that P≠NP, we do not agree that this weighs against the plausibility of intelligence explosion. (Note that due to a miscommunication between McDermott and the editors, a faulty draft of McDermott [2012a] was published in *Journal of Consciousness Studies*. We recommend reading the corrected version at `http://cs-www.cs.yale.edu/homes/dvm/papers/chalmers-singularity-response.pdf`.)

have the same optimization power, but one seems to be optimizing more intelligently. For this reason, we adopt a description of intelligence as optimization power divided by resources used (Yudkowsky 2008b).[3] For our purposes, "intelligence" measures an agent's capacity for *efficient* cross-domain optimization of the world according to the agent's preferences. Using this definition, we can avoid common objections to the use of human-centric notions of intelligence in discussions of the technological singularity (Greenfield 2012), and hopefully we can avoid common anthropomorphisms that often arise when discussing intelligence (Muehlhauser and Helm 2012).

By "AI," we refer to general AI rather than narrow AI. That is, we refer to "systems which match or exceed the [intelligence] of humans in virtually all domains of interest" (Shulman and Bostrom 2012). By this definition, IBM's *Jeopardy!*-playing computer Watson is not an "AI" (in our sense) but merely a *narrow* AI, because it can only solve a narrow set of problems. Drop Watson in a pond or ask it to do original science, and it would be helpless even if given a month's warning to prepare. Imagine instead a machine that could invent new technologies, manipulate humans with acquired social skills, and otherwise learn to navigate many new social and physical environments as needed to achieve its goals.

Which kinds of machines might accomplish such feats? There are many possible types. A *whole brain emulation* (WBE) would be a computer emulation of brain structures sufficient to functionally reproduce human cognition. We need not understand the mechanisms of general intelligence to use the human intelligence software already invented by evolution (Sandberg and Bostrom 2008). In contrast, "*de novo* AI" requires inventing intelligence software anew. There is a vast space of possible mind designs for *de novo* AI (Dennett 1996; Yudkowsky 2008a). *De novo* AI approaches include the symbolic, probabilistic, connectionist, evolutionary, embedded, and other research programs (Pennachin and Goertzel 2007).

## 2. From Here to AI

When should we expect the first creation of AI? We must allow for a wide range of possibilities. Except for weather forecasters (Murphy and Winkler 1984), and successful professional gamblers, nearly all of us give inaccurate probability estimates, and in particular we are overconfident of our predictions (Lichtenstein, Fischhoff, and Phillips

3. This definition is a useful starting point, but it could be improved. Future work could produce a definition of intelligence as optimization power over a canonical distribution of environments, with a penalty for resource use—e.g. the "speed prior" described by Schmidhuber (2002). Also see Goertzel (2006, 48; 2010) and Hibbard (2011).

1982; Griffin and Tversky 1992; Yates et al. 2002). This overconfidence affects professional forecasters, too (Tetlock 2005), and we have little reason to think AI forecasters have fared any better.[4] So if you have a gut feeling about when AI will be created, it is probably wrong.

But uncertainty is not a "get out of prediction free" card (Bostrom 2007). We still need to decide whether or not to encourage WBE development, whether or not to help fund AI safety research, etc. Deciding either way already implies some sort of prediction. Choosing not to fund AI safety research suggests that we do not think AI is near, while funding AI safety research implies that we think AI might be coming soon.

### 2.1. Predicting AI

How, then, might we predict when AI will be created? We consider several strategies below.

*By gathering the wisdom of experts or crowds.* Many experts and groups have tried to predict the creation of AI. Unfortunately, experts' predictions are often little better than those of laypeople (Tetlock 2005), expert elicitation methods have in general not proven useful for long-term forecasting,[5] and prediction markets (ostensibly drawing on the opinions of those who believe themselves to possess some expertise) have not yet been demonstrated useful for technological forecasting (Williams 2011). Still, it may be useful to note that none to few experts expect AI within five years, whereas many experts expect AI by 2050 or 2100.[6]

*By simple hardware extrapolation.* The novelist Vernor Vinge (1993) based his own predictions about AI on hardware trends, but in a 2003 reprint of his article, Vinge notes the insufficiency of this reasoning: even if we acquire hardware sufficient for AI, we may not have the software problem solved.[7]

Hardware extrapolation may be a more useful method in a context where the intelligence software is already written: whole brain emulation. Because WBE seems to rely

4. To take one of many examples, Simon (1965, 96) predicted that "machines will be capable, within twenty years, of doing any work a man can do." Also see Crevier (1993).

5. Armstrong (1985), Woudenberg (1991), and Rowe and Wright (2001). But, see Parente and Anderson-Parente (2011).

6. Bostrom (2003), Bainbridge and Roco (2006), Legg (2008), Baum, Goertzel, and Goertzel (2011), Sandberg and Bostrom (2011), and Nielsen (2011).

7. A software bottleneck may delay AI but create greater risk. If there is a software bottleneck on AI, then when AI is created there may be a "computing overhang": large amounts of inexpensive computing power which could be used to run thousands of AIs or give a few AIs vast computational resources. This may not be the case if early AIs require quantum computing hardware, which is less likely to be plentiful and inexpensive than classical computing hardware at any given time.

mostly on scaling up existing technologies like microscopy and large-scale cortical simulation, WBE may be largely an "engineering" problem, and thus the time of its arrival may be more predictable than is the case for other kinds of AI.

Several authors have discussed the difficulty of WBE in detail (Kurzweil 2005; Sandberg and Bostrom 2008; de Garis et al. 2010; Modha et al. 2011; Cattell and Parker 2012). In short: The difficulty of WBE depends on many factors, and in particular on the resolution of emulation required for successful WBE. For example, proteome-resolution emulation would require more resources and technological development than emulation at the resolution of the brain's neural network. In perhaps the most likely scenario,

> WBE on the neuronal/synaptic level requires relatively modest increases in microscopy resolution, a less trivial development of automation for scanning and image processing, a research push at the problem of inferring functional properties of neurons and synapses, and relatively business-as-usual development of computational neuroscience models and computer hardware. (Sandberg and Bostrom 2008, 83)

*By considering the time since Dartmouth.* We have now seen more than 50 years of work toward machine intelligence since the seminal Dartmouth conference on AI, but AI has not yet arrived. This seems, intuitively, like strong evidence that AI won't arrive in the next minute, good evidence it won't arrive in the next year, and significant but far from airtight evidence that it won't arrive in the next few decades. Such intuitions can be formalized into models that, while simplistic, can form a useful starting point for estimating the time to machine intelligence.[8]

8. We can make a simple formal model of this evidence by assuming (with much simplification) that every year a coin is tossed to determine whether we will get AI that year, and that we are initially unsure of the weighting on that coin. We have observed more than 50 years of "no AI" since the first time serious scientists believed AI might be around the corner. This "56 years of no AI" observation would be highly unlikely under models where the coin comes up "AI" on 90% of years (the probability of our observations would be $10^{-56}$), or even models where it comes up "AI" in 10% of all years (probability 0.3%), whereas it's the expected case if the coin comes up "AI" in, say, 1% of all years, or for that matter in 0.0001% of all years. Thus, in this toy model, our "no AI for 56 years" observation should update us strongly against coin weightings in which AI would be likely in the next minute, or even year, while leaving the relative probabilities of "AI expected in 200 years" and "AI expected in 2 million years" more or less untouched. (These updated probabilities are robust to choice of the time interval between coin flips; it matters little whether the coin is tossed once per decade, or once per millisecond, or whether one takes a limit as the time interval goes to zero.) Of course, one gets a different result if a different "starting point" is chosen, e.g. Alan Turing's seminal paper on machine intelligence (Turing 1950), or the inaugural conference on artificial general intelligence (Wang, Goertzel, and Franklin 2008). For more on this approach and Laplace's rule of succession, see Jaynes (2003, chap. 18). We suggest this approach

*By tracking progress in machine intelligence.* Some people intuitively estimate the time until AI by asking what proportion of human abilities today's software can match, and how quickly machines are catching up.[9] However, it is not clear how to divide up the space of "human abilities," nor how much each one matters. We also don't know if progress in machine intelligence will be linear, exponential, or otherwise. Watching an infant's progress in learning calculus might lead one to infer the child will not learn it until the year 3000, until suddenly the child learns it in a spurt at age 17. Still, it may be worth asking whether a measure can be found for which both: (a) progress is predictable enough to extrapolate; and (b) when performance rises to a certain level, we can expect AI.

*By extrapolating from evolution.* Evolution managed to create intelligence without using intelligence to do so. Perhaps this fact can help us establish an upper bound on the difficulty of creating AI (Chalmers 2010; Moravec 1976, 1998, 1999), though this approach is complicated by observation selection effects (Shulman and Bostrom 2012).

*By estimating progress in scientific research output.* Imagine a man digging a ten-kilometer ditch. If he digs 100 meters in one day, you might predict the ditch will be finished in 100 days. But what if 20 more diggers join him, and they are all given backhoes? Now the ditch might not take so long. Analogously, when predicting progress toward AI it may be useful to consider not how much progress is made per year, but instead how much progress is made per unit of research effort, and how many units of research effort we can expect to be applied to the problem in the coming decades.

Unfortunately, we have not yet discovered demonstrably reliable methods for long-term technological forecasting. New methods are being tried (Nagy et al. 2010), but until they prove successful we should be particularly cautious when predicting AI timelines. Below, we attempt a final approach by examining some plausible *speed bumps* and *accelerators* on the path to AI.

### 2.2. Speed Bumps

Several factors may decelerate our progress toward the first creation of AI. For example:

*An end to Moore's law.* Though several information technologies have progressed at an exponential or superexponential rate for many decades (Nagy et al. 2011), this trend may not hold for much longer (Mack 2011).

---

only as a way of generating a prior probability distribution over AI timelines, from which one can then update upon encountering additional evidence.

9. Relatedly, Good (1970) tried to predict the first creation of AI by surveying past conceptual breakthroughs in AI and extrapolating into the future.

*Depletion of low-hanging fruit.* Scientific progress is not only a function of research effort but also of the ease of scientific discovery; in some fields there is pattern of increasing difficulty with each successive discovery (Arbesman 2011; Jones 2009). AI may prove to be a field in which new discoveries require far more effort than earlier discoveries.

*Societal collapse.* Various political, economic, technological, or natural disasters may lead to a societal collapse during which scientific progress would not continue (Posner 2004; Bostrom and Ćirković 2008).

*Disinclination.* Chalmers (2010) and Hutter (2012) think the most likely speed bump in our progress toward AI will be disinclination, including active prevention. Perhaps humans will not want to create their own successors. New technologies like "Nanny AI" (Goertzel 2012), or new political alliances like a stable global totalitarianism (Caplan 2008), may empower humans to delay or prevent scientific progress that could lead to the creation of AI.

### 2.3. Accelerators

Other factors, however, may accelerate progress toward AI:

*More hardware.* For at least four decades, computing power[10] has increased exponentially, roughly in accordance with Moore's law.[11] Experts disagree on how much longer Moore's law will hold (Mack 2011; Lundstrom 2003), but even if hardware advances more slowly than exponentially, we can expect hardware to be far more powerful in a few decades than it is now.[12] More hardware doesn't by itself give us machine intelligence, but it contributes to the development of machine intelligence in several ways:

> Powerful hardware may improve performance simply by allowing existing "brute force" solutions to run faster (Moravec 1976). Where such solutions do not yet exist, researchers might be incentivized to quickly develop them given abundant hardware to exploit. Cheap computing may enable much more extensive experimentation in algorithm design, tweaking parameters or using methods such as genetic algorithms. Indirectly, computing may enable

10. The technical measure predicted by Moore's law is the density of components on an integrated circuit, but this is closely tied to the price-performance of computing power.

11. For important qualifications, see Nagy et al. (2010) and Mack (2011).

12. Quantum computing may also emerge during this period. Early worries that quantum computing may not be feasible have been overcome, but it is hard to predict whether quantum computing will contribute significantly to the development of machine intelligence because progress in quantum computing depends heavily on relatively unpredictable insights in quantum algorithms and hardware (Rieffel and Polak 2011).

> the production and processing of enormous datasets to improve AI performance (Halevy, Norvig, and Pereira 2009), or result in an expansion of the information technology industry and the quantity of researchers in the field. (Shulman and Sandberg 2010)

*Better algorithms.* Often, mathematical insights can reduce the computation time of a program by many orders of magnitude without additional hardware. For example, IBM's Deep Blue played chess at the level of world champion Garry Kasparov in 1997 using about 1.5 trillion instructions per second (TIPS), but a program called Deep Junior did it in 2003 using only 0.015 TIPS. Thus, the computational efficiency of the chess algorithms increased by a factor of 100 in only six years (Richards and Shaw 2004).

*Massive datasets.* The greatest leaps forward in speech recognition and translation software have come not from faster hardware or smarter hand-coded algorithms, but from access to massive data sets of human-transcribed and human-translated words (Halevy, Norvig, and Pereira 2009). Datasets are expected to increase greatly in size in the coming decades, and several technologies promise to actually *outpace* "Kryder's law" (Kryder and Kim 2009), which states that magnetic disk storage density doubles approximately every 18 months (Walter 2005).

*Progress in psychology and neuroscience.* Cognitive scientists have uncovered many of the brain's algorithms that contribute to human intelligence (Trappenberg 2009; Ashby and Helie 2011). Methods like neural networks (imported from neuroscience) and reinforcement learning (inspired by behaviorist psychology) have already resulted in significant AI progress, and experts expect this insight-transfer from neuroscience to AI to continue and perhaps accelerate (Van der Velde 2010; Schierwagen 2011; Floreano and Mattiussi 2008; de Garis et al. 2010; Krichmar and Wagatsuma 2011).

*Accelerated science.* A growing First World will mean that more researchers at well-funded universities will be conducting research relevant to machine intelligence. The world's scientific output (in publications) grew by one third from 2002 to 2007 alone, much of this driven by the rapid growth of scientific output in developing nations like China and India (Royal Society 2011).[13] Moreover, new tools can accelerate particular fields, just as fMRI accelerated neuroscience in the 1990s, and the effectiveness of scientists themselves can potentially be increased with cognitive enhancement pharmaceuticals (Bostrom and Sandberg 2009), and brain-computer interfaces that allow direct neural access to large databases (Groß 2009). Finally, new collaborative tools like blogs

13. On the other hand, some worry (Pan et al. 2005), that the rates of scientific fraud and publication bias may currently be higher in China and India than in the developed world.

and Google Scholar are already yielding results such as the Polymath Project, which is rapidly and collaboratively solving open problems in mathematics (Nielsen 2011).[14]

*Economic incentive.* As the capacities of "narrow AI" programs approach the capacities of humans in more domains (Koza 2010), there will be increasing demand to replace human workers with cheaper, more reliable machine workers (Hanson 2008, 1998; Kaas et al. 2010; Brynjolfsson and McAfee 2011).

*First-mover incentives.* Once AI looks to be within reach, political and private actors will see substantial advantages in building AI first. AI could make a small group more powerful than the traditional superpowers—a case of "bringing a gun to a knife fight." The race to AI may even be a "winner take all" scenario. Thus, political and private actors who realize that AI is within reach may devote substantial resources to developing AI as quickly as possible, provoking an AI arms race (Gubrud 1997).

### 2.4. How Long, Then, Before AI?

So, when will we create AI? Any predictions on the matter must have wide error bars. Given the history of confident false predictions about AI (Crevier 1993), and AI's potential speed bumps, it seems misguided to be 90% confident that AI will succeed in the coming century. But 90% confidence that AI will *not* arrive before the end of the century also seems wrong, given that: (a) many difficult AI breakthroughs have now been made, (b) several factors, such as automated science and first-mover incentives, may well accelerate progress toward AI, and (c) whole brain emulation seems to be possible and have a more predictable development than *de novo* AI. Thus, we think there is a significant probability that AI will be created this century. This claim is not scientific—the field of technological forecasting is not yet advanced enough for that—but we believe our claim is reasonable.

The creation of human-level AI would have serious repercussions, such as the displacement of most or all human workers (Brynjolfsson and McAfee 2011). But if AI is likely to lead to machine superintelligence, as we argue next, the implications could be even greater.

14. Also, a process called "iterated embryo selection" (Uncertain Future 2012), could be used to produce an entire generation of scientists with the cognitive capabilities of Albert Einstein or John von Neumann, thus accelerating scientific progress and giving a competitive advantage to nations which choose to make use of this possibility.

## 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.

Thinking that FAI may be too difficult, Goertzel (2012) proposes a global "Nanny AI" that would "forestall a full-on Singularity for a while, . . . giving us time to figure out what kind of Singularity we really want to build and how." Goertzel and others working on AI safety theory would very much appreciate the extra time to solve the problems of AI safety before the first self-improving AI is created, but your authors suspect that Nanny AI is "FAI-complete," or nearly so. That is, in order to build Nanny AI, you may need to solve all the problems required to build full-blown Friendly AI, for example the problem of specifying precise goals (Yudkowsky 2011; Muehlhauser and Helm 2012), and the problem of maintaining a stable utility function under radical self-modification, including updates to the AI's internal ontology (de Blanc 2011).

The approaches to controlled intelligence explosion we have surveyed so far attempt to constrain an AI's goals, but others have suggested a variety of "external" constraints for goal-directed AIs: physical and software confinement (Chalmers 2010; Yampolskiy 2012), deterrence mechanisms, and tripwires that shut down an AI if it engages in dangerous behavior. Unfortunately, these solutions would pit human intelligence against superhuman intelligence, and we shouldn't be confident the former would prevail.

Perhaps we could build an AI of limited cognitive ability—say, a machine that only answers questions: an "Oracle AI." But this approach is not without its own dangers (Armstrong, Sandberg, and Bostrom 2012).

Unfortunately, even if these latter approaches worked, they might merely delay AI risk without eliminating it. If one AI development team has successfully built either an Oracle AI or a goal-directed AI under successful external constraints, other AI development teams may not be far from building their own AIs, some of them with less effective safety measures. A Friendly AI with enough lead time, however, could permanently prevent the creation of unsafe AIs.

### 4.2. What Can We Do About AI Risk?

Because superhuman AI and other powerful technologies may pose some risk of human extinction ("existential risk"), Bostrom (2002) recommends a program of *differential technological development* in which we would attempt "to retard the implementation of dangerous technologies and accelerate implementation of beneficial technologies, especially those that ameliorate the hazards posed by other technologies."

But good outcomes from intelligence explosion appear to depend not only on differential technological development but also, for example, on solving certain kinds of problems in decision theory and value theory before the first creation of AI (Muehlhauser 2011). Thus, we recommend a course of *differential intellectual progress*, which includes differential technological development as a special case.

Differential intellectual progress consists in prioritizing risk-*reducing* intellectual progress over risk-*increasing* intellectual progress. As applied to AI risks in particular, a plan of differential intellectual progress would recommend that our progress on the scientific, philosophical, and technological problems of AI *safety* outpace our progress on the problems of AI *capability* such that we develop *safe* superhuman AIs before we develop (arbitrary) superhuman AIs. Our first superhuman AI must be a safe superhuman AI, for we may not get a second chance (Yudkowsky 2008a). With AI as with other technologies, we may become victims of "the tendency of technological advance to outpace the social control of technology" (Posner 2004).

## 5. Conclusion

We have argued that AI poses an existential threat to humanity. On the other hand, with more intelligence we can hope for quicker, better solutions to many of our problems. We don't usually associate cancer cures or economic stability with artificial intelligence, but curing cancer is ultimately a problem of being smart enough to figure out how to cure it, and achieving economic stability is ultimately a problem of being smart enough to figure out how to achieve it. To whatever extent we have goals, we have goals that can be accomplished to greater degrees using sufficiently advanced intelligence. When considering the likely consequences of superhuman AI, we must respect both risk and opportunity.[26]

26. Our thanks to Nick Bostrom, Steve Rayhawk, David Chalmers, Steve Omohundro, Marcus Hutter, Brian Rabkin, William Naaktgeboren, Michael Anissimov, Carl Shulman, Eliezer Yudkowsky, Louie Helm, Jesse Liptrap, Nisan Stiennon, Will Newsome, Kaj Sotala, Julia Galef, and anonymous reviewers for their helpful comments.

## References


Anderson, Britt. 1993. "Evidence from the Rat for a General Factor That Underlies Cognitive Performance and That Relates to Brain Size: Intelligence?" *Neuroscience Letters* 153 (1): 98–102. doi:10.1016/0304-3940(93)90086-Z.

Arbesman, Samuel. 2011. "Quantifying the Ease of Scientific Discovery." *Scientometrics* 86 (2): 245–250. doi:10.1007/s11192-010-0232-6.

Armstrong, Jon Scott. 1985. *Long-Range Forecasting: From Crystal Ball to Computer.* 2nd ed. New York: Wiley.

Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. 2012. "Thinking Inside the Box: Controlling and Using an Oracle AI." *Minds and Machines* 22 (4): 299–324. doi:10.1007/s11023-012-9282-2.

Ashby, F. Gregory, and Sebastien Helie. 2011. "A Tutorial on Computational Cognitive Neuroscience: Modeling the Neurodynamics of Cognition." *Journal of Mathematical Psychology* 55 (4): 273–289. doi:10.1016/j.jmp.2011.04.003.

Bainbridge, William Sims, and Mihail C. Roco, eds. 2006. *Managing Nano-Bio-Info-Cogno Innovations: Converging Technologies in Society.* Dordrecht, The Netherlands: Springer.

Baum, Seth D., Ben Goertzel, and Ted G. Goertzel. 2011. "How Long Until Human-Level AI? Results from an Expert Assessment." *Technological Forecasting and Social Change* 78 (1): 185–195. doi:10.1016/j.techfore.2010.09.006.

Block, Ned. 1981. "Psychologism and Behaviorism." *Philosophical Review* 90 (1): 5–43. doi:10.2307/2184371.

Bostrom, Nick. Forthcoming. *Superintelligence: A Strategic Analysis of the Coming Machine Intelligence Revolution.* Manuscript, in preparation.

———. 2002. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9. http://www.jetpress.org/volume9/risks.html.

———. 2003. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence,* edited by Iva Smit and George E. Lasker, 2:12–17. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

———. 2006. "What is a Singleton?" *Linguistic and Philosophical Investigations* 5 (2): 48–54.

———. 2007. "Technological Revolutions: Ethics and Policy in the Dark." In *Nanoscale: Issues and Perspectives for the Nano Century,* edited by Nigel M. de S. Cameron and M. Ellen Mitchell, 129–152. Hoboken, NJ: John Wiley & Sons. doi:10.1002/9780470165874.ch10.

———. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." In "Theory and Philosophy of AI," edited by Vincent C. Müller, special issue, *Minds and Machines* 22 (2): 71–85. doi:10.1007/s11023-012-9281-3.

Bostrom, Nick, and Milan M. Ćirković, eds. 2008. *Global Catastrophic Risks.* New York: Oxford University Press.

Bostrom, Nick, and Anders Sandberg. 2009. "Cognitive Enhancement: Methods, Ethics, Regulatory Challenges." *Science and Engineering Ethics* 15 (3): 311–341. doi:10.1007/s11948-009-9142-5.

Brynjolfsson, Erik, and Andrew McAfee. 2011. *Race Against The Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy.* Lexington, MA: Digital Frontier. Kindle edition.

Caplan, Bryan. 2008. "The Totalitarian Threat." In Bostrom and Ćirković 2008, 504–519.

Cartwright, Edward. 2011. *Behavioral Economics.* Routledge Advanced Texts in Economics and Finance. New York: Routledge.

Cattell, Rick, and Alice Parker. 2012. *Challenges for Brain Emulation: Why is Building a Brain so Difficult?* Synaptic Link, February 5. http://synapticlink.org/Brain%20Emulation%20Challenges.pdf.

Chalmers, David John. 1996. *The Conscious Mind: In Search of a Fundamental Theory.* Philosophy of Mind Series. New York: Oxford University Press.

———. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (9–10): 7–65. http://www.ingentaconnect.com/content/imp/jcs/2010/00000017/f0020009/art00001.

———. 2012. "The Singularity: A Reply to Commentators." *Journal of Consciousness Studies* 19 (7-8): 141–167. http://ingentaconnect.com/content/imp/jcs/2012/00000019/F0020007/art00014.

Crevier, Daniel. 1993. *AI: The Tumultuous History of the Search for Artificial Intelligence.* New York: Basic Books.

De Blanc, Peter. 2011. *Ontological Crises in Artificial Agents' Value Systems.* The Singularity Institute, San Francisco, CA, May 19. http://arxiv.org/abs/1105.3821.

De Garis, Hugo, Chen Shuo, Ben Goertzel, and Lian Ruiting. 2010. "A World Survey of Artificial Brain Projects, Part I: Large-Scale Brain Simulations." *Neurocomputing* 74 (1–3): 3–29. doi:10.1016/j.neucom.2010.08.004.

Dennett, Daniel C. 1996. *Kinds of Minds: Toward an Understanding of Consciousness.* Science Master. New York: Basic Books.

Dreyfus, Hubert L. 1972. *What Computers Can't Do: A Critique of Artificial Reason.* New York: Harper & Row.

Eden, Amnon, Johnny Søraker, James H. Moor, and Eric Steinhart, eds. 2012. *Singularity Hypotheses: A Scientific and Philosophical Assessment.* The Frontiers Collection. Berlin: Springer.

Feldman, J. A., and Dana H. Ballard. 1982. "Connectionist Models and Their Properties." *Cognitive Science* 6 (3): 205–254. doi:10.1207/s15516709cog0603_1.

Floreano, Dario, and Claudio Mattiussi. 2008. *Bio-Inspired Artificial Intelligence: Theories, Methods, and Technologies.* Intelligent Robotics and Autonomous Agents. Cambridge, MA: MIT Press.

Fox, Douglas. 2011. "The Limits of Intelligence." *Scientific American,* July, 36–43.

Fregni, Felipe, Paulo S. Boggio, Michael Nitsche, Felix Bermpohl, Andrea Antal, Eva Feredoes, Marco A. Marcolin, et al. 2005. "Anodal Transcranial Direct Current Stimulation of Prefrontal Cortex Enhances Working Memory." *Experimental Brain Research* 166 (1): 23–30. doi:10.1007/s00221-005-2334-6.

Friedman, James W., ed. 1994. *Problems of Coordination in Economic Activity.* Recent Economic Thought 35. Boston: Kluwer Academic.

Friedman, Milton. 1953. "The Methodology of Positive Economics." In *Essays in Positive Economics,* 3–43. Chicago: University of Chicago Press.

Goertzel, Ben. 2006. *The Hidden Pattern: A Patternist Philosophy of Mind.* Boca Raton, FL: BrownWalker.

———. 2010. "Toward a Formal Characterization of Real-World General Intelligence." In *Artificial General Intelligence: Proceedings of the Third Conference on Artificial General Intelligence, AGI 2010, Lugano, Switzerland, March 5–8, 2010,* edited by Eric B. Baum, Marcus Hutter, and Emanuel Kitzelmann, 19–24. Advances in Intelligent Systems Research 10. Amsterdam: Atlantis. doi:10.2991/agi.2010.17.

———. 2012. "Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood?" *Journal of Consciousness Studies* 19 (1–2): 96–111. http://ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00006.

Goertzel, Ben, and Cassio Pennachin, eds. 2007. *Artificial General Intelligence.* Cognitive Technologies. Berlin: Springer. doi:10.1007/978-3-540-68677-4.

Goldreich, Oded. 2010. *P, NP, and NP-Completeness: The Basics of Computational Complexity.* New York: Cambridge University Press.

Good, Irving John. 1959. *Speculations on Perceptrons and Other Automata.* Research Lecture, RC-115. IBM, Yorktown Heights, New York, June 2. http://domino.research.ibm.com/library/cyberdig.nsf/papers/58DC4EA36A143C218525785E00502E30/$File/rc115.pdf.

———. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers,* edited by Franz L. Alt and Morris Rubinoff, 6:31–88. New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0.

———. 1970. "Some Future Social Repercussions of Computers." *International Journal of Environmental Studies* 1 (1–4): 67–79. doi:10.1080/00207237008709398.

———. 1982. "Ethical Machines." In *Intelligent Systems: Practice and Perspective,* edited by J. E. Hayes, Donald Michie, and Y.-H. Pao, 555–560. Machine Intelligence 10. Chichester: Ellis Horwood.

Greenfield, Susan. 2012. "The Singularity: Commentary on David Chalmers." *Journal of Consciousness Studies* 19 (1–2): 112–118. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00007.

Griffin, Dale, and Amos Tversky. 1992. "The Weighing of Evidence and the Determinants of Confidence." *Cognitive Psychology* 24 (3): 411–435. doi:10.1016/0010-0285(92)90013-R.

Groß, Dominik. 2009. "Blessing or Curse? Neurocognitive Enhancement by 'Brain Engineering.'" *Medicine Studies* 1 (4): 379–391. doi:10.1007/s12376-009-0032-6.

Gubrud, Mark Avrum. 1997. "Nanotechnology and International Security." Paper presented at the Fifth Foresight Conference on Molecular Nanotechnology, Palo Alto, CA, November 5–8. http://www.foresight.org/Conferences/MNT05/Papers/Gubrud/.

Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. "The Unreasonable Effectiveness of Data." *IEEE Intelligent Systems* 24 (2): 8–12. doi:10.1109/MIS.2009.36.

Hanson, Robin. 1998. "Economic Growth Given Machine Intelligence." Unpublished manuscript. Accessed May 15, 2013. http://hanson.gmu.edu/aigrow.pdf.

———. 2008. "Economics of the Singularity." *IEEE Spectrum* 45 (6): 45–50. doi:10.1109/MSPEC.2008.4531461.

———. 2012. "Meet the New Conflict, Same as the Old Conflict." *Journal of Consciousness Studies* 19 (1–2): 119–125. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00008.

Hanson, Robin, and Eliezer Yudkowsky. 2008. "The Hanson-Yudkowsky AI-Foom Debate." Less Wrong Wiki. Accessed March 13, 2012. http://wiki.lesswrong.com/wiki/The_Hanson-Yudkowsky_AI-Foom_Debate.

Heylighen, Francis. 2012. "Brain in a Vat Cannot Break Out." *Journal of Consciousness Studies* 19 (1–2): 126–142. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00009.

Hibbard, Bill. 2011. "Measuring Agent Intelligence via Hierarchies of Environments." In Schmidhuber, Thórisson, and Looks 2011, 303–308.

Hutter, Marcus. 2005. *Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability.* Texts in Theoretical Computer Science. Berlin: Springer. doi:10.1007/b138233.

———. 2012. "Can Intelligence Explode?" *Journal of Consciousness Studies* 19 (1–2): 143–166. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00010.

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science.* Edited by G. Larry Bretthorst. New York: Cambridge University Press. doi:10.2277/0521592712.

Jones, Benjamin F. 2009. "The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?" *Review of Economic Studies* 76 (1): 283–317. doi:10.1111/j.1467-937X.2008.00531.x.

Kaas, Steven, Steve Rayhawk, Anna Salamon, and Peter Salamon. 2010. *Economic Implications of Software Minds.* The Singularity Institute, San Francisco, CA, August 10. http://intelligence.org/files/EconomicImplications.pdf.

Kandel, Eric R., James H. Schwartz, and Thomas M. Jessell, eds. 2000. *Principles of Neural Science.* New York: McGraw-Hill.

Koza, John R. 2010. "Human-Competitive Results Produced by Genetic Programming." *Genetic Programming and Evolvable Machines* 11 (3–4): 251–284. doi:10.1007/s10710-010-9112-3.

Krichmar, Jeffrey L., and Hiroaki Wagatsuma, eds. 2011. *Neuromorphic and Brain-Based Robots.* New York: Cambridge University Press.

Kryder, M. H., and Chang Soo Kim. 2009. "After Hard Drives—What Comes Next?" *IEEE Transactions on Magnetics* 2009 (10): 3406–3413. doi:10.1109/TMAG.2009.2024163.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology.* New York: Viking.

Legg, Shane. 2008. "Machine Super Intelligence." PhD diss., University of Lugano. http://www.vetta.org/documents/Machine_Super_Intelligence.pdf.

Legg, Shane, and Marcus Hutter. 2007. "A Collection of Definitions of Intelligence." In *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms—Proceedings of the AGI Workshop 2006,* edited by Ben Goertzel and Pei Wang, 17–24. Frontiers in Artificial Intelligence and Applications 157. Amsterdam: IOS.

Lichtenstein, Sarah, Baruch Fischhoff, and Lawrence D. Phillips. 1982. "Calibration of Probabilities: The State of the Art to 1980." In *Judgement Under Uncertainty: Heuristics and Biases,* edited by Daniel Kahneman, Paul Slovic, and Amos Tversky, 306–334. New York: Cambridge University Press.

Loosemore, Richard, and Ben Goertzel. 2011. "Why an Intelligence Explosion is Probable." *H+ Magazine,* March 7. http://hplusmagazine.com/2011/03/07/why-an-intelligence-explosion-is-probable/.

Lucas, J. R. 1961. "Minds, Machines and Gödel." *Philosophy* 36 (137): 112–127. doi:10.1017/S0031819100057983.

Lundstrom, Mark. 2003. "Moore's Law Forever?" *Science* 299 (5604): 210–211. doi:10.1126/science.1079567.

Mack, C. A. 2011. "Fifty Years of Moore's Law." *IEEE Transactions on Semiconductor Manufacturing* 24 (2): 202–207. doi:10.1109/TSM.2010.2096437.

Marcus, Gary. 2008. *Kluge: The Haphazard Evolution of the Human Mind.* Boston: Houghton Mifflin.

McAfee, Andrew, and Erik Brynjolfsson. 2008. "Investing in the IT That Makes a Competitive Difference." *Harvard Business Review,* July. http://hbr.org/2008/07/investing-in-the-it-that-makes-a-competitive-difference.

McCorduck, Pamela. 2004. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence.* 2nd ed. Natick, MA: A K Peters.

McDaniel, Michael A. 2005. "Big-Brained People are Smarter: A Meta-Analysis of the Relationship between In Vivo Brain Volume and Intelligence." *Intelligence* 33 (4): 337–346. doi:10.1016/j.intell.2004.11.005.

McDermott, Drew. 2012a. "Response to 'The Singularity' by David Chalmers." *Journal of Consciousness Studies* 19 (1–2): 167–172. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00011.

———. 2012b. "There Are No 'Extendible Methods' in David Chalmers's Sense Unless P=NP." Unpublished manuscript. Accessed March 19, 2012. http://cs-www.cs.yale.edu/homes/dvm/papers/no-extendible-methods.pdf.

Mehta, Ghanshyam B. 1998. "Preference and Utility." In *Handbook of Utility Theory, Volume I,* edited by Salvador Barbera, Peter J. Hammond, and Christian Seidl, 1–47. Boston: Kluwer Academic.

Minsky, Marvin. 1984. "Afterword to Vernor Vinge's novel, 'True Names.'" Unpublished manuscript, October 1. Accessed December 31, 2012. http://web.media.mit.edu/~minsky/papers/TrueNames.Afterword.html.

Modha, Dharmendra S., Rajagopal Ananthanarayanan, Steven K. Esser, Anthony Ndirango, Anthony J. Sherbondy, and Raghavendra Singh. 2011. "Cognitive Computing." *Communications of the ACM* 54 (8): 62–71. doi:10.1145/1978542.1978559.

Modis, Theodore. 2012. "There Will Be No Singularity." In Eden, Søraker, Moor, and Steinhart 2012.

Moravec, Hans P. 1976. "The Role of Raw Rower in Intelligence." Unpublished manuscript, May 12. Accessed August 12, 2012. http://www.frc.ri.cmu.edu/users/hpm/project.archive/general.articles/1975/Raw.Power.html.

———. 1998. "When Will Computer Hardware Match the Human Brain?" *Journal of Evolution and Technology* 1. http://www.transhumanist.com/volume1/moravec.htm.

———. 1999. "Rise of the Robots." *Scientific American,* December, 124–135.

Muehlhauser, Luke. 2011. "So You Want to Save the World." Last revised March 2, 2012. http://lukeprog.com/SaveTheWorld.html.

Muehlhauser, Luke, and Louie Helm. 2012. "The Singularity and Machine Ethics." In Eden, Søraker, Moor, and Steinhart 2012.

Murphy, Allan H., and Robert L. Winkler. 1984. "Probability Forecasting in Meteorology." *Journal of the American Statistical Association* 79 (387): 489–500.

Nagy, Béla, J. Doyne Farmer, Jessika E. Trancik, and Quan Minh Bui. 2010. *Testing Laws of Technological Progress.* Santa Fe, NM: Santa Fe Institute, September 2. http://tuvalu.santafe.edu/~bn/workingpapers/NagyFarmerTrancikBui.pdf.

Nagy, Béla, J. Doyne Farmer, Jessika E. Trancik, and John Paul Gonzales. 2011. "Superexponential Long-Term Trends in Information Technology." *Technological Forecasting and Social Change* 78 (8): 1356–1364. doi:10.1016/j.techfore.2011.07.006.

Nielsen, Michael. 2011. "What Should a Reasonable Person Believe about the Singularity?" *Michael Nielsen* (blog), January 12. http://michaelnielsen.org/blog/what-should-a-reasonable-person-believe-about-the-singularity/.

Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements.* New York: Cambridge University Press.

Nordmann, Alfred. 2007. "If and Then: A Critique of Speculative NanoEthics." *NanoEthics* 1 (1): 31–46. doi:10.1007/s11569-007-0007-6.

Omohundro, Stephen M. 1987. "Efficient Algorithms with Neural Network Behavior." *Complex Systems* 1 (2): 273–347. http://www.complex-systems.com/abstracts/v01_i02_a04.html.

———. 2007. "The Nature of Self-Improving Artificial Intelligence." Paper presented at Singularity Summit 2007, San Francisco, CA, September 8–9. http://selfawaresystems.com/2007/10/05/paper-on-the-nature-of-self-improving-artificial-intelligence/.

———. 2008. "The Basic AI Drives." In Wang, Goertzel, and Franklin 2008, 483–492.

———. 2012. "Rational Artificial Intelligence for the Greater Good." In Eden, Søraker, Moor, and Steinhart 2012.

Pan, Zhenglun, Thomas A. Trikalinos, Fotini K. Kavvoura, Joseph Lau, and John P. A. Ioannidis. 2005. "Local Literature Bias in Genetic Epidemiology: An Empirical Evaluation of the Chinese Literature." *PLoS Medicine* 2 (12): e334. doi:10.1371/journal.pmed.0020334.

Parente, Rick, and Janet Anderson-Parente. 2011. "A Case Study of Long-Term Delphi Accuracy." *Technological Forecasting and Social Change* 78 (9): 1705–1711. doi:10.1016/j.techfore.2011.07.005.

Pennachin, Cassio, and Ben Goertzel. 2007. "Contemporary Approaches to Artificial General Intelligence." In Goertzel and Pennachin 2007, 1–30.

Penrose, Roger. 1994. *Shadows of the Mind: A Search for the Missing Science of Consciousness.* New York: Oxford University Press.

Plebe, Alessio, and Pietro Perconti. 2012. "The Slowdown Hypothesis." In Eden, Søraker, Moor, and Steinhart 2012.

Posner, Richard A. 2004. *Catastrophe: Risk and Response.* New York: Oxford University Press.

Proudfoot, Diane, and B. Jack Copeland. 2012. "Artificial Intelligence." In *The Oxford Handbook of Philosophy of Cognitive Science,* edited by Eric Margolis, Richard Samuels, and Stephen P. Stich. New York: Oxford University Press.

Richards, Mark A., and Gary A. Shaw. 2004. "Chips, Architectures and Algorithms: Reflections on the Exponential Growth of Digital Signal Processing Capability." Unpublished manuscript, January 28. Accessed March 20, 2012. http://users.ece.gatech.edu/~mrichard/Richards&Shaw_Algorithms01204.pdf.

Rieffel, Eleanor, and Wolfgang Polak. 2011. *Quantum Computing: A Gentle Introduction.* Scientific and Engineering Computation. Cambridge, MA: MIT Press.

Rowe, Gene, and George Wright. 2001. "Expert Opinions in Forecasting: The Role of the Delphi Technique." In *Principles of Forecasting: A Handbook for Researchers and Practitioners,* edited by Jon Scott Armstrong. International Series in Operations Research & Management Science 30. Boston: Kluwer Academic.

The Royal Society. 2011. *Knowledge, Networks and Nations: Global Scientific Collaboration in the 21st Century.* RS Policy document, 03/11. Royal Society, London. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/publications/2011/4294976134.pdf.

Russell, Stuart J., and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach.* 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Sandberg, Anders. 2010. "An Overview of Models of Technological Singularity." Paper presented at the Roadmaps to AGI and the Future of AGI Workshop, Lugano, Switzerland, March 8. http://agi-conf.org/2010/wp-content/uploads/2009/06/agi10singmodels2.pdf.

———. 2011. "Cognition Enhancement: Upgrading the Brain." In *Enhancing Human Capacities,* edited by Julian Savulescu, Ruud ter Meulen, and Guy Kahane, 71–91. Malden, MA: Wiley-Blackwell.

Sandberg, Anders, and Nick Bostrom. 2008. *Whole Brain Emulation: A Roadmap.* Technical Report, 2008-3. Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/wp-content/uploads/brain-emulation-roadmap-report1.pdf.

———. 2011. *Machine Intelligence Survey.* Technical Report, 2011-1. Future of Humanity Institute, University of Oxford. www.fhi.ox.ac.uk/reports/2011-1.pdf.

Schaul, Tom, and Jürgen Schmidhuber. 2010. "Metalearning." *Scholarpedia* 5 (6): 4650. doi:10.4249/scholarpedia.4650.

Schierwagen, Andreas. 2011. "Reverse Engineering for Biologically Inspired Cognitive Architectures: A Critical Analysis." In *From Brains to Systems: Brain-Inspired Cognitive Systems 2010,* edited by Carlos Hernández, Ricardo Sanz, Jaime Gómez-Ramirez, Leslie S. Smith, Amir Hussain, Antonio Chella, and Igor Aleksander, 111–121. Advances in Experimental Medicine and Biology 718. New York: Springer. doi:10.1007/978-1-4614-0164-3_10.

Schmidhuber, Jürgen. 2002. "The Speed Prior: A New Simplicity Measure Yielding Near-Optimal Computable Predictions." In *Computational Learning Theory: 5th Annual Conference on Computational Learning Theory, COLT 2002 Sydney, Australia, July 8–10, 2002 Proceedings,* edited by Jyrki Kivinen and Robert H. Sloan, 123–127. Lecture Notes in Computer Science 2375. Berlin: Springer. doi:10.1007/3-540-45435-7_15.

———. 2007. "Gödel Machines: Fully Self-Referential Optimal Universal Self-Improvers." In Goertzel and Pennachin 2007, 199–226.

Schmidhuber, Jürgen, Kristinn R. Thórisson, and Moshe Looks, eds. 2011. *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings.* Lecture Notes in Computer Science 6830. Berlin: Springer. doi:10.1007/978-3-642-22887-2.

Schneider, Stefan. 2010. *Homo Economicus—or More like Homer Simpson?* Current Issues. Deutsche Bank Research, Frankfurt, June 29. http://www.dbresearch.com/PROD/DBR_INTERNET_EN-PROD/PROD0000000000259291.PDF.

Schoenemann, Paul Thomas. 1997. "An MRI Study of the Relationship Between Human Neuroanatomy and Behavioral Ability." PhD diss., University of California, Berkeley. http://mypage.iu.edu/~toms/papers/dissertation/Dissertation_title.htm.

Schwartz, Jacob T. 1987. "Limits of Artificial Intelligence." In *Encyclopedia of Artificial Intelligence,* edited by Stuart C. Shapiro and David Eckroth, 1:488–503. New York: John Wiley & Sons.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (03): 417–424. doi:10.1017/S0140525X00005756.

Shulman, Carl, and Nick Bostrom. 2012. "How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects." *Journal of Consciousness Studies* 19 (7–8): 103–130. http://ingentaconnect.com/content/imp/jcs/2012/00000019/F0020007/art00011.

Shulman, Carl, and Anders Sandberg. 2010. "Implications of a Software-Limited Singularity." In *ECAP10: VIII European Conference on Computing and Philosophy,* edited by Klaus Mainzer. Munich: Dr. Hut.

Simon, Herbert Alexander. 1965. *The Shape of Automation for Men and Management.* New York: Harper & Row.

Solomonoff, Ray J. 1985. "The Time Scale of Artificial Intelligence: Reflections on Social Effects." *Human Systems Management* 5:149–153.

Sotala, Kaj. 2012. "Advantages of Artificial Intelligences, Uploads, and Digital Minds." *International Journal of Machine Consciousness* 4 (1): 275–291. doi:10.1142/S1793843012400161.

Tetlock, Philip E. 2005. *Expert Political Judgment: How Good is it? How Can We Know?* Princeton, NJ: Princeton University Press.

Trappenberg, Thomas P. 2009. *Fundamentals of Computational Neuroscience.* 2nd ed. New York: Oxford University Press.

Turing, A. M. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236): 433–460. doi:10.1093/mind/LIX.236.433.

———. 1951. "Intelligent Machinery, A Heretical Theory." A lecture given to '51 Society' at Manchester.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124–1131. doi:10.1126/science.185.4157.1124.

———. 1983. "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment." *Psychological Review* 90 (4): 293–315. doi:10.1037/0033-295X.90.4.293.

The Uncertain Future. 2012. "What is Multi-generational In Vitro Embryo Selection?" The Uncertain Future. Accessed March 25, 2012. http://www.theuncertainfuture.com/faq.html#7.

Van der Velde, Frank. 2010. "Where Artificial Intelligence and Neuroscience Meet: The Search for Grounded Architectures of Cognition." *Advances in Artificial Intelligence,* no. 5. doi:10.1155/2010/918062.

Van Gelder, Timothy, and Robert F. Port. 1995. "It's About Time: An Overview of the Dynamical Approach to Cognition." In *Mind as Motion: Explorations in the Dynamics of Cognition,* edited by Robert F. Port and Timothy van Gelder. Bradford Books. Cambridge, MA: MIT Press.

Vinge, Vernor. 1993. "The Coming Technological Singularity: How to Survive in the Post-Human Era." In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace,* 11–22. NASA Conference Publication 10129. NASA Lewis Research Center. http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855_1994022855.pdf.

Von Neumann, John. 1966. *Theory of Self-Replicating Automata.* Edited by Arthur Walter Burks. Urbana: University of Illinois Press.

Walter, Chip. 2005. "Kryder's Law." *Scientific American,* July 25. http://www.scientificamerican.com/article.cfm?id=kryders-law.

Wang, Pei, Ben Goertzel, and Stan Franklin, eds. 2008. *Artificial General Intelligence 2008: Proceedings of the First AGI Conference.* Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

Williams, Leighton Vaughan, ed. 2011. *Prediction Markets: Theory and Applications.* Routledge International Studies in Money and Banking 66. New York: Routledge.

Wootters, W. K., and W. H. Zurek. 1982. "A Single Quantum Cannot Be Cloned." *Nature* 299 (5886): 802–803. doi:10.1038/299802a0.

Woudenberg, Fred. 1991. "An Evaluation of Delphi." *Technological Forecasting and Social Change* 40 (2): 131–150. doi:10.1016/0040-1625(91)90002-W.

Yampolskiy, Roman V. 2012. "Leakproofing the Singularity: Artificial Intelligence Confinement Problem." *Journal of Consciousness Studies* 2012 (1–2): 194–214. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00014.

Yates, J. Frank, Ju-Whei Lee, Winston R. Sieck, Incheol Choi, and Paul C. Price. 2002. "Probability Judgment Across Cultures." In *Heuristics and Biases: The Psychology of Intuitive Judgment,* edited by Thomas Gilovich, Dale Griffin, and Daniel Kahneman, 271–291. New York: Cambridge University Press. doi:10.2277/0521796792.

Yudkowsky, Eliezer. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures.* The Singularity Institute, San Francisco, CA, June 15. http://intelligence.org/files/CFAI.pdf.

———. 2008a. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In Bostrom and Ćirković 2008, 308–345.

———. 2008b. "Efficient Cross-Domain Optimization." *Less Wrong* (blog), October 28. http://lesswrong.com/lw/vb/efficient_crossdomain_optimization/.

———. 2011. "Complex Value Systems in Friendly AI." In Schmidhuber, Thórisson, and Looks 2011, 388–393.