Chapter 2
Intelligence Explosion: Evidence and Import

Luke Muehlhauser and Anna Salamon

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.

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

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.

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

L. Muehlhauser (✉) · A. Salamon
Machine Intelligence Research Institute, Berkeley, CA 94705, USA
e-mail: luke@singularity.org

A. H. Eden et al. (eds.), *Singularity Hypotheses*, The Frontiers Collection,
DOI: 10.1007/978-3-642-32560-1_2, © Springer-Verlag Berlin Heidelberg 2012

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 2012a); 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, forthcoming) and Bostrom (Forthcoming[a]). 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 (see Modis, this volume), nor indeed that AI progress will accelerate rather than decelerate (see Plebe and Perconti, this volume). 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, Chaps. 8 and 9; Legg 2008; Heylighen 2012) or to the possibility of intelligence explosion (Chalmers, forthcoming).[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 have the same optimization power, but one seems to be optimizing more intelligently. For this reason, we adopt Yudkowsky's (2008b) description of intelligence as optimization power divided by resources used.[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, this volume).

---

[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, b) 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 From AI to Machine Superintelligence 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 (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.).

[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, p. 48, 2010), Hibbard (2011).

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).

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 et al. 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 ave 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.

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

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 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 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, p. 83)

---

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

[6] Bostrom (2003), Bainbridge (2006), Legg (2008), Baum et al. (2011), Sandberg and Bostrom (2011), 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.

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]

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 10 km 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

[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 et al. 2008). For more on this approach and Laplace's rule of succession, see Jaynes (2003), Chap. 18. We suggest this approach 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.

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.

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).

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), Hutter (2012a) 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.

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

[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), Mack (2011).

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 the production and processing of enormous datasets to improve AI performance (Halevi et al. 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 (Halevi et al. 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

[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).

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 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, Forthcoming; 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).

How Long, Then, Before AI?

We have not yet mentioned two small but significant developments leading us to agree with Schmidhuber (2012) that "progress towards self-improving AIs is already substantially beyond what many futurists and philosophers are aware of." These two developments are Marcus Hutter's universal and provably optimal AIXI agent model (Hutter 2005) and Jürgen Schmidhuber's universal self-improving Gödel machine models (Schmidhuber 2007, 2009).

Schmidhuber (2012) summarizes the importance of the Gödel machine:

> [The] Gödel machine… already *is* a universal AI that is at least theoretically optimal in certain sense. It may interact with some initially unknown, partially observable environment to maximize future expected utility or reward by solving arbitrary user-defined computational tasks. Its initial algorithm is not hardwired; it can completely rewrite itself without essential limits apart from the limits of computability, provided a proof searcher

[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.

[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.

> embedded within the initial algorithm can first prove that the rewrite is useful, according to the formalized utility function taking into account the limited computational resources. Self-rewrites may modify/improve the proof searcher itself, and can be shown to be *globally optimal*, relative to Gödel's well-known fundamental restrictions of provability (Gödel 1931)…
>
> All of this implies that there already exists the blueprint of a Universal AI which will solve almost all problems almost as quickly as if it already knew the best (unknown) algorithm for solving them, because almost all imaginable problems are big enough to make the additive constant negligible. Hence, I must object to Chalmers' statement [that] "we have yet to find the right algorithms, and no-one has come close to finding them yet."

Next, we turn to Hutter (2012b) for a summary of the importance of AIXI:

> The concrete ingredients in AIXI are as follows: Intelligent *actions* are based on informed *decisions*. Attaining good decisions requires *predictions* which are typically based on models of the environments. Models are constructed or learned from past observations via *induction*. Fortunately, based on the *deep philosophical insights* and *powerful mathematical developments*, all these problems have been overcome, at least in theory: So what do we need (from a mathematical point of view) to construct a universal optimal learning agent interacting with an arbitrary unknown environment? The theory, coined *UAI* [Universal Artificial Intelligence], developed in the last decade and explained in Hutter (2005) says: *All you need is Ockham, Epicurus, Turing, Bayes, Solomonoff* (1964a, 1964b)*, Kolmogorov* (1968)*, Bellman* (1957): Sequential decision theory (Bertsekas 2007) (*Bellman's* equation) formally solves the problem of rational agents in uncertain worlds if the true environmental probability distribution is known. If the environment is unknown, Bayesians (Berger 1993) replace the true distribution by a weighted mixture of distributions from some (hypothesis) class. Using the large class of all (semi)measures that are (semi)computable on a *Turing* machine bears in mind *Epicurus*, who teaches not to discard any (consistent) hypothesis. In order not to ignore *Ockham*, who would select the simplest hypothesis, *Solomonoff* defined a universal prior that assigns high/low prior weight to simple/complex environments (Rathmanner and Hutter 2011), where *Kolmogorov* quantifies complexity (Li and Vitányi 2008). Their unification constitutes the theory of UAI and resulted in… AIXI.[15]

AIXI is incomputable, but computationally tractable approximations have already been experimentally tested, and these reveal a path to universal AI[16] that solves real-world problems in a variety of environments:

[15] In our two quotes from Hutter (2012b) we have replaced Hutter's AMS-style citations with Chicago-style citations.

[16] The creation of AI probably is not, however, merely a matter of finding computationally tractable AIXI approximations that can solve increasingly complicated problems in increasingly complicated environments. There remain many open problems in the theory of universal artificial intelligence (Hutter 2009). For problems related to allowing some AIXI-like models to self-modify, see Orseau and Ring (2011), Ring and Orseau (2011), Orseau (2011); Hibbard (Forthcoming). Dewey (2011) explains why reinforcement learning agents like AIXI may pose a threat to humanity.

> The same single [AIXI approximation "MC-AIXI-CTW"] is already able to learn to play TicTacToe, Kuhn Poker, and most impressively Pacman (Veness et al. 2011) from scratch. Besides Pacman, there are hundreds of other arcade games from the 1980s, and it would be sensational if a single algorithm could learn them all solely by trial and error, which seems feasible for (a variant of) MC-AIXI-CTW. While these are "just" recreational games, they do contain many prototypical elements of the real world, such as food, enemies, friends, space, obstacles, objects, and weapons. Next could be a test in modern virtual worlds… that require intelligent agents, and finally some selected real-world problems.

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 (including the Gödel machine and AIXI), (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.

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 superintelligence 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.

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).[17]

Communication speed. Axons carry spike signals at 75 meters per second or less (Kandel et al. 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.[18]

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

[17] 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".

[18] For example see Omohundro (1987).

current state can be copied,[19] 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.[20]

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 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.[21] 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 (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 (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 (Stanovich 2010; Cartwright

[19] 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 (Wooters and Zurek 1982).

[20] 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).

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

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; Russel and Norvig 2009, Chap. 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.

Instrumentally Convergent Goals

Omohundro (2007, 2008, this volume) and Bostrom (Forthcoming[a]) 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.[22]
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.

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

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.[23] 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 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[a]), and also about whether intelligence explosion would result in a stable equilibrium of multiple machine superintelligence or instead a machine "singleton" (Bostrom 2006). We will not discuss these complex issues here.

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).

[23] 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, forthcoming). 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.

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.[24]

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 (including 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."

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?[25] WBEs and other brain-inspired AIs running on human-derived "spaghetti code" (Marcus 2008) may not have a clear "slot" in which to specify desirable goals. 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,[26] 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, this volume), and allowing an AI to *learn* desirable goals

[24] This may be less true for early-generation WBEs, but Omohundro (2008) argues that AIs will converge upon being optimizing agents, which exhibit a strict division between goals and cognitive ability.

[25] 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 (Forthcoming).

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

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).[27]

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."

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, this volume) 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 Forthcoming; Bostrom forthcoming).

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.

[27] 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.

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).

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.[28]

[28] 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, B. (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, S. (2011). Quantifying the ease of scientific discovery. *Scientometrics, 86*(2), 245–250. doi:10.1007/s11192-010-0232-6.

Armstrong, J. S. (1985). *Long-range forecasting: from crystal ball to computer* (2nd ed.). New York: Wiley.

Armstrong, S., Sandberg, A., & Bostrom N. Forthcoming. Thinking inside the box: using and controlling an Oracle AI. *Minds and Machines.*

Ashby, F. G., & Helie S. (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, W. S., & Roco, M. C. (Eds.). (2006). *Managing nano-bio-info-cogno innovations: converging technologies in society*. Dordrecht: Springer.

Baum, S. D., Goertzel, B., & Goertzel, T. G. (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.

Bellman, R. E. (1957). *Dynamic programming*. Princeton: Princeton University Press.

Berger, J. O. (1993). Statistical decision theory and bayesian analysis (2nd edn). *Springer Series in Statistics*. New York: Springer.

Bertsekas, D. P. (2007). *Dynamic programming and optimal control* (Vol. 2). Nashua: Athena Scientific.

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

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology, 9* http://www.jetpress.org/volume9/risks.html.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. In I. Smit & G. E. Lasker (Eds.), *Cognitive, emotive and ethical aspects of decision making in humans and in artificial intelligence*. Windsor: International Institute of Advanced Studies in Systems Research/ Cybernetics. Vol. 2.

Bostrom, N. (2006). What is a singleton? *Linguistic and Philosophical Investigations, 5*(2), 48–54.

Bostrom, N. (2007). Technological revolutions: Ethics and policy in the dark. In M. Nigel, S. de Cameron, & M. E. Mitchell (Eds.), *Nanoscale: Issues and perspectives for the nano century* (pp. 129–152). Hoboken: Wiley. doi:10.1002/9780470165874.ch10.

Bostrom, N. Forthcoming(a). *Superintelligence: A strategic analysis of the coming machine intelligence revolution.* Manuscript, in preparation.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines.* Preprint at, http://www.nickbostrom.com/superintelligentwill.pdf.

Bostrom, N., & Ćirković, M. M. (Eds.). (Eds.). (2008). *Global catastrophic risks*. New York: Oxford University Press.

Bostrom, N., & Sandberg, A. (2009). Cognitive enhancement: Methods, ethics, regulatory challenges. *Science and Engineering Ethics, 15*(3), 311–341. doi:10.1007/s11948-009-9142-5.

Brynjolfsson, E., & McAfee, A. (2011). *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*. Lexington: Digital Frontier Press. Kindle edition.

Caplan, B. (2008). The totalitarian threat. In Bostrom and Ćirković 2008, 504–519.

Cartwright, E. (2011). *Behavioral economics*. New York: Routledge Advanced Texts in Economics and Finance.

Cattell, R, & Parker, A. (2012). *Challenges for brain emulation: why is building a brain so difficult?* Synaptic Link, Feb. 5. http://synapticlink.org/Brain%20Emulation%20 Challenges.pdf.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. New York: Oxford University Press. (Philosophy of Mind Series).

Chalmers, D. J. (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.

Chalmers, D. J. Forthcoming. The singularity: A reply. *Journal of Consciousness Studies 19*.

Crevier, D. (1993). *AI: The tumultuous history of the search for artificial intelligence*. New York: Basic Books.

de Blanc, P. (2011). *Ontological crises in artificial agents' value systems.* San Francisco: Singularity Institute for Artificial Intelligence, May 19. http://arxiv.org/abs/1105.3821.

de Garis, H., Shuo, C., Goertzel, B., & Ruiting, L. (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, D. C. (1996). *Kinds of minds: Toward an understanding of consciousness*., Science Master New York: Basic Books.

Dewey, D. (2011). Learning what to value. In Schmidhuber, J., Thórisson, KR., & Looks, M. 2011, 309–314.

Dreyfus, H. L. (1972). *What computers can't do: A critique of artificial reason*. New York: Harper & Row.

Eden, A., Søraker, J., Moor, J. H., & Steinhart, E. (Eds.). (2012). *The singularity hypothesis: A scientific and philosophical assessment*. Berlin: Springer.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science, 6*(3), 205–254. doi:10.1207/s15516709cog0603_1.

Floreano, D., & Mattiussi, C. (2008). *Bio-inspired artificial intelligence: Theories, methods, and technologies.* Intelligent Robotics and Autonomous Agents. MIT Press: Cambridge.

Fox, D. (2011). The limits of intelligence. Scientific American, July, 36–43.

Fregni, F., Boggio, P. S., Nitsche, M., Bermpohl, F., Antal, A., Feredoes, E., 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, M. (1953). The methodology of positive economics. In *Essays in positive economics* (pp. 3–43). Chicago: Chicago University Press.

Friedman, James W., (Ed.) (1994). *Problems of coordination in economic activity* (Vol. 35). Recent Economic Thought. Boston: Kluwer Academic Publishers.

Gödel, K. (1931). Über formal unentscheidbare sätze der Principia Mathematica und verwandter systeme I. *Monatshefte für Mathematik, 38*(1), 173–198. doi:10.1007/BF01700692.

Goertzel, B. (2006). *The hidden pattern: A patternist philosophy of mind*. Boco Raton: BrownWalker Press.

Goertzel, B. (2010). Toward a formal characterization of real-world general intelligence. In E. Baum, M. Hutter, & E. Kitzelmann (Eds.) *Artificial general intelligence: Proceedings of the third conference on artificial general intelligence, AGI 2010, Lugano, Switzerland, March 5–8, 2010*, 19–24. Vol. 10. Advances in Intelligent Systems Research. Amsterdam: Atlantis Press. doi:10.2991/agi.2010.17.

Goertzel, B. (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, B., & Pennachin, C. (Eds.) (2007). *Artificial general intelligence. Cognitive Technologies.* Berlin: Springer. doi:10.1007/978-3-540-68677-4.

Goldreich, O. (2010). *P, NP, and NP-Completeness: The basics of computational complexity*. New York: Cambridge University Press.

Good, I. J. (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.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. In F. L. Alt & M. Rubinoff (Eds.) *Advances in computers* (pp. 31–88. Vol. 6). New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0.

Good, I. J. (1970). Some future social repercussions of computers. *International Journal of Environmental Studies, 1*(1–4), 67–79. doi:10.1080/00207237008709398.

Good, I. J. (1982). Ethical machines. In J. E. Hayes, D. Michie, & Y.-H. Pao (Eds.) *Machine intelligence* (pp. 555–560, Vol. 10). Intelligent Systems: Practice and Perspective. Chichester: Ellis Horwood.

Greenfield, S. (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, D., & Tversky, A. (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, D. (2009). Blessing or curse? Neurocognitive enhancement by "brain engineering". *Medicine Studies, 1*(4), 379–391. doi:10.1007/s12376-009-0032-6.

Gubrud, M. A. (1997). Nanotechnology and international security. Paper presented at the Fifth Foresight Conference on Molecular Nanotechnology, Palo Alto, CA, Nov. 5–8. http://www.foresight.org/Conferences/MNT05/Papers/Gubrud/.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8–12. doi:10.1109/MIS.2009.36.

Hanson, R. (2008). Economics of the singularity. *IEEE Spectrum, 45*(6), 45–50. doi:10.1109/MSPEC.2008.4531461.

Hanson, R. (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, R. Forthcoming. Economic growth given machine intelligence. *Journal of Artificial Intelligence Research.*

Hanson, R., & Yudkowsky, E. (2008). The Hanson-Yudkowsky AI-foom debate. LessWrong Wiki. http://wiki.lesswrong.com/wiki/The_Hanson-Yudkowsky_AI-Foom_Debate (accessed Mar. 13, 2012).

Hibbard, B. (2011). Measuring agent intelligence via hierarchies of environments. In Schmidhuber, J., Thórisson, KR., & Looks, M. 2011, 303–308.

Hibbard, B. Forthcoming. Model-based utility functions. *Journal of Artificial General Intelligence.*

Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Texts in Theoretical Computer Science. Berlin: Springer. doi:10.1007/b138233.

Hutter, M. (2009). Open problems in universal induction & intelligence. *Algorithms, 2*(3), 879–906. doi:10.3390/a2030879.

Hutter, M. (2012a). Can intelligence explode? *Journal of Consciousness Studies 19*(1–2), 143–166. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00010.

Hutter, M. (2012b). One decade of universal artificial intelligence. In P. Wang & B. Goertzel (eds.) *Theoretical foundations of artificial general intelligence* (Vol. 4). Atlantis Thinking Machines. Paris: Atlantis Press.

Jaynes, E. T., & Bretthorst, G. L. (Eds.) (2003). *Probability theory: The logic of science*. New York: Cambridge University Press. doi:10.2277/0521592712.

Jones, B. 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, S., Rayhawk,S., Salamon, A., & Salamon, P. (2010). *Economic implications of software minds.* San Francisco: Singularity Institute for Artificial Intelligence, Aug. 10. http://www.singinst.co/upload/economic-implications.pdf.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (Eds.). (2000). *Principles of neural science*. New York: McGraw-Hill.

Kolmogorov, A. N. (1968). Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics, 2*(1–4), 157–168. doi:10.1080/00207166808803030.

Koza, J. 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, J. L., & Wagatsuma, H. (Eds.). (2011). *Neuromorphic and brain-based robots*. New York: Cambridge University Press.

Kryder, M. H., & Kim, C. S. (2009). After hard drives—what comes next? *IEEE Transactions on Magnetics, 2009*(10), 3406–3413. doi:10.1109/TMAG.2009.2024163.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. New York: Viking.

Lampson, B. W. (1973). A note on the confinement problem. *Communications of the ACM, 16*(10), 613–615. doi:10.1145/362375.362389.

Legg, S. (2008). Machine super intelligence. PhD diss., University of Lugano. http://www.vetta.org/documents/Machine_Super_Intelligence.pdf.

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. In B. Goertzel & P. Wang (Eds.) *Advances in artificial general intelligence: Concepts, architectures and algorithms—proceedings of the AGI workshop 2006* (Vol. 157). Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press.

Li, M., & Vitányi, P. M. B. (2008). An introduction to Kolmogorov complexity and its applications. Texts in Computer Science. New York: Springer. doi:10.1007/978-0-387-49820-1.

Lichtenstein, S., Fischoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press.

Loosmore, R., & Goertzel, B. (2011). Why an intelligence explosion is probable. *H+ Magazine,* Mar. 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, M. (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, G. (2008). *Kluge: The haphazard evolution of the human mind*. Boston: Houghton Mifflin.

McAfee, A., & Brynjolfsson, E. (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, P. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence* (2nd ed.). Natick: A. K. Peters.

McDaniel, M. 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, D. (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.

McDermott, D. (2012b). There are no "Extendible Methods" in David Chalmers's sense unless P=NP. Unpublished manuscript. http://cs-www.cs.yale.edu/homes/dvm/papers/no-extendible-methods.pdf (accessed Mar. 19, 2012).

Mehta, G. B. (1998). Preference and utility. In S. Barbera, P. J. Hammond, & C. Seidl (Eds.), *Handbook of utility theory* (Vol. I, pp. 1–47). Boston: Kluwer Academic Publishers.

Minsky, M. (1984). Afterword to Vernor Vinge's novel, "True Names." Unpublished manuscript, Oct. 1. http://web.media.mit.edu/~minsky/papers/TrueNames.Afterword.html (accessed Mar. 26, 2012).

Modha, D. S., Ananthanarayanan, R., Esser, S. K., Ndirango, A., Sherbondy, A. J., & Singh, R. (2011). Cognitive computing. *Communications of the ACM, 54*(8), 62–71. doi:10.1145/1978542.1978559.

Modis, T. (2012). There will be no singularity. In Eden, Søraker, Moor, & Steinhart 2012.

Moravec, H. P. (1976). The role of raw rower in intelligence. May 12. http://www.frc.ri.cmu.edu/users/hpm/project.archive/general.articles/1975/Raw.Power.html (accessed Mar. 13, 2012).

Moravec, H. (1998). When will computer hardware match the human brain? *Journal of Evolution and Technology* 1. http://www.transhumanist.com/volume1/moravec.htm.

Moravec, H. (1999). Rise of the robots. *Scientific American,* Dec., 124–135.

Muehlhauser, L. (2011). So you want to save the world. Last modified Mar. 2, 2012. http://lukeprog.com/SaveTheWorld.html.

Muehlhauser, L., & Helm, L. (2012). The singularity and machine ethics. In Eden, Søraker, Moor, & Steinhart 2012.

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association, 79*(387), 489–500.

Nagy, B., Farmer, J. D., Trancik, J. E., & Bui, QM. (2010). *Testing laws of technological progress.* Santa Fe Institute, NM, Sept. 2. http://tuvalu.santafe.edu/ bn/workingpapers/NagyFarmerTrancikBui.pdf.

Nagy, B., Farmer, J. D., Trancik, J. E., & Gonzales, J. P. (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, M. (2011). What should a reasonable person believe about the singularity? Michael Nielsen (blog). Jan. 12. http://michaelnielsen.org/blog/what-should-a-reasonable-person-believe-about-the-singularity/ (accessed Mar. 13, 2012).

Nilsson, N. J. (2009). *The quest for artificial intelligence: A history of ideas and achievements*. New York: Cambridge University Press.

Nordmann, A. (2007). If and then: A critique of speculative nanoethics. *NanoEthics, 1*(1), 31–46. doi:10.1007/s11569-007-0007-6.

Omohundro, S. M. (1987). Efficient algorithms with neural network behavior. *Complex Systems 1*(2), 273–347. http://www.complex-systems.com/abstracts/v01_i02_a04.html.

Omohundro, S. M. (2007). The nature of self-improving artificial intelligence. Paper presented at the Singularity Summit 2007, San Francisco, CA, Sept. 8–9. http://singinst.org/summit2007/overview/abstracts/#omohundro.

Omohundro, S. M. (2008). The basic AI drives. In Wang, Goertzel, & Franklin 2008, 483–492.

Omohundro, S. M. 2012. Rational artificial intelligence for the greater good. In Eden, Søraker, Moor, & Steinhart 2012.

Orseau, L. (2011). Universal knowledge-seeking agents. In *Algorithmic learning theory: 22nd international conference, ALT 2011, Espoo, Finland, October 5–7, 2011. Proceedings,* ed. Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann. Vol. 6925. Lecture Notes in Computer Science. Berlin: Springer. doi:10.1007/978-3-642-24412-4_28.

Orseau, L., & Ring, M. (2011). Self-modification and mortality in artificial agents. In Schmidhuber, Thórisson, and Looks 2011, 1–10.

Pan, Z., Trikalinos, T. A., Kavvoura, F. K., Lau, J., & Ioannidis, J. P. A. (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, R., & Anderson-Parente, J. (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, C, & Goertzel, B. (2007). Contemporary approaches to artificial general intelligence. In Goertzel & Pennachin 2007, 1–30.

Penrose, R. (1994). *Shadows of the mind: A search for the missing science of consciousness*. New York: Oxford University Press.

Plebe, A., & Perconti, P. (2012). The slowdown hypothesis. In Eden, Søraker, Moor, & Steinhart 2012.

Posner, R. A. (2004). *Catastrophe: Risk and response*. New York: Oxford University Press.

Proudfoot, D., & Jack Copeland, B. (2012). Artificial intelligence. In E. Margolis, R. Samuels, & S. P. Stich (Eds.), *The Oxford handbook of philosophy of cognitive science*. New York: Oxford University Press.

Rathmanner, S., & Hutter, M. (2011). A philosophical treatise of universal induction. *Entropy, 13*(6), 1076–1136. doi:10.3390/e13061076.

Richards, M. A., & Shaw, G. A. (2004). Chips, architectures and algorithms: Reflections on the exponential growth of digital signal processing capability. Unpublished manuscript, Jan. 28. http://users.ece.gatech.edu/ mrichard/Richards&Shaw_Algorithms01204.pdf (accessed Mar. 20, 2012).

Rieffel, E., & Polak, W. (2011). *Quantum computing: A gentle introduction.* Scientific and Engineering Computation. Cambridge: MIT Press.

Ring, M., & Orseau, L. (2011). Delusion, survival, and intelligent agents. In Schmidhuber, Thórisson, & Looks 2011, 11–20.

Rowe, G., & Wright, G. (2001). Expert opinions in forecasting: The role of the Delphi technique. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners,* (Vol. 30). International Series in Operations Research & Management Science. Boston: Kluwer Academic Publishers.

Russell, S. J., & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd ed.). Upper Saddle River: Prentice-Hall.

Sandberg, A. (2010). An overview of models of technological singularity. Paper presented at the Roadmaps to AGI and the future of AGI workshop, Lugano, Switzerland, Mar. 8th. http://agi-conf.org/2010/wp-content/uploads/2009/06/agi10singmodels2.pdf.

Sandberg, A. (2011). Cognition enhancement: Upgrading the brain. In J. Savulescu, R. ter Meulen, & G. Kahane (Eds.), *Enhancing human capacities* (pp. 71–91). Malden: Wiley-Blackwell.

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

Sandberg, A., & Bostrom, N. (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, T., & Schmidhuber, J. (2010). Metalearning. *Scholarpedia, 5*(6), 4650. doi:10.4249/scholarpedia.4650.

Schierwagen, A. (2011). Reverse engineering for biologically inspired cognitive architectures: A critical analysis. In C. Hernández, R. Sanz, J. Gómez-Ramirez, L. S. Smith, A. Hussain, A. Chella, & I. Aleksander (Eds.), *From brains to systems: Brain-inspired cognitive systems 2010,* (pp. 111–121, Vol. 718). Advances in Experimental Medicine and Biology. New York: Springer. doi:10.1007/978-1-4614-0164-3_10.

Schmidhuber, J. (2002). The speed prior: A new simplicity measure yielding near-optimal computable predictions. In J. Kivinen & R. H. Sloan, *Computational learning theory: 5th annual conference on computational learning theory, COLT 2002 Sydney, Australia, July 8–10, 2002 proceedings,* (pp. 123–127, Vol. 2375). Lecture Notes in Computer Science. Berlin: Springer. doi:10.1007/3-540-45435-7_15.

Schmidhuber, J. (2007). Gödel machines: Fully self-referential optimal universal self-improvers. In Goertzel & Pennachin 2007, 199–226.

Schmidhuber, J. (2009). Ultimate cognition *à la* Gödel. *Cognitive Computation, 1*(2), 177–193. doi:10.1007/s12559-009-9014-y.

Schmidhuber, J. (2012). Philosophers & futurists, catch up! Response to The Singularity. *Journal of Consciousness Studies 19*(1–2), 173–182. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00012.

Schmidhuber, J., Thórisson, K. R., & Looks, M. (Eds.) (2011). *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings* (Vol. 6830). Lecture Notes in Computer Science. Berlin: Springer. doi:10.1007/978-3-642-22887-2.

Schneider, S. (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, P. T. (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, J. T. (1987). Limits of artificial intelligence. In S. C. Shapiro & D. Eckroth (Eds.), *Encyclopedia of artificial intelligence* (pp. 488–503, Vol. 1). New York: Wiley.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(03), 417–424. doi:10.1017/S0140525X00005756.

Shulman, C., & Bostrom, N. (2012). How hard is artificial intelligence? Evolutionary arguments and selection effects. *Journal of Consciousness Studies 19*.

Shulman, C., & Sandberg, A. (2010). Implications of a software-limited singularity. Paper presented at the 8th European Conference on Computing and Philosophy (ECAP), Munich, Germany, Oct. 4–6.

Simon, H. A. (1965). *The shape of automation for men and management*. New York: Harper & Row.

Solomonoff, R. J. (1964a). A formal theory of inductive inference. *Part I. Information and Control, 7*(1), 1–22. doi:10.1016/S0019-9958(64)90223-2.

Solomonoff, R. J. (1964b). A formal theory of inductive inference. *Part II. Information and Control, 7*(2), 224–254. doi:10.1016/S0019-9958(64)90131-7.

Solomonoff, R. J. (1985). The time scale of artificial intelligence: Reflections on social effects. *Human Systems Management, 5*, 149–153.

Sotala, K. (2012). Advantages of artificial intelligences, uploads, and digital minds. *International Journal of Machine Consciousness 4*.

Stanovich, K. E. (2010). *Rationality and the reflective mind*. New York: Oxford University Press.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?*. Princeton: Princeton University Press.

The Royal Society. (2011). *Knowledge, networks and nations: Global scientific collaboration in the 21st century.* RS Policy document, 03/11. The Royal Society, London. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/publications/2011/4294976134.pdf.

Trappenberg, T. 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.

Turing, A. M. (1951). Intelligent machinery, a heretical theory. A lecture given to '51 Society' at Manchester.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. doi:10.1126/science.185.4157.1124.

Tversky, A., & Kahneman, D. (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. http://www.theuncertainfuture.com/faq.html#7 (accessed Mar. 25, 2012).

Van der Velde, F. (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, T., & Port, R. F. (1995). It's about time: An overview of the dynamical approach to cognition. In R. F. Port & T. van Gelder. *Mind as motion: Explorations in the dynamics of cognition,* Bradford Books. Cambridge: MIT Press.

Veness, J., Ng, K. S., Hutter, M., Uther, W., & Silver, D. (2011). A Monte-Carlo AIXI approximation. *Journal of Artificial Intelligence Research, 40*, 95–142. doi:10.1613/jair.3125.

Vinge, V. (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, J., & Burks, A. W. (Eds.) (1966). *Theory of self-replicating automata.* Urbana: University of Illinois Press.

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

Wang, P., Goertzel, B., & Franklin, S. (Eds.). (2008). *Artificial General Intelligence 2008: Proceedings of the First AGI Conference (Vol. 171). Frontiers in Artificial Intelligence and Applications*. Amsterdam: IOS Press.

Williams, L. V. (Ed.). (2011). *Prediction markets: Theory and applications (Vol. 66). Routledge International Studies in Money and Banking*. New York: Routledge.

Wootters, W. K., & Zurek, W. H. (1982). A single quantum cannot be cloned. *Nature, 299*(5886), 802–803. doi:10.1038/299802a0.

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

Yampolskiy, R. V. (2012). Leakproofing the singularity: Artificial intelligence confinement problem. *Journal of Consciousness Studies 19*(1–2), 194–214. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00014.

Yates, J. F., Lee, J.-W., Sieck, W. R., Choi, I., & Price, P. C. (2002). Probability judgment across cultures. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 271–291). New York: Cambridge University Press.

Yudkowsky, E. (2001). Creating Friendly AI 1.0: The analysis and design of benevolent goal architectures. The Singularity Institute, San Francisco, CA, June 15. http://singinst.org/upload/CFAI.html.

Yudkowsky, E. (2008a). Artificial intelligence as a positive and negative factor in global risk. In Bostrom & Ćirković 2008, 308–345.

Yudkowsky, E. (2008b). Efficient cross-domain optimization. LessWrong. Oct. 28. http://lesswrong.com/lw/vb/efficient_crossdomain_optimization/ (accessed Mar. 19, 2012).

Yudkowsky, E. (2011). Complex value systems in friendly AI. In Schmidhuber, Thórisson, & Looks 2011, 388–393.

Chapter 2A
Robin Hanson on Muehlhauser and Salamon's "Intelligence Explosion: Evidence and Import"

Muehlhauser and Salamon [M&S] talk as if their concerns are particular to an unprecedented new situation: the imminent prospect of "artificial intelligence" (AI). But in fact their concerns depend little on how artificial will be our descendants, nor on how intelligence they will be. Rather, Muehlhauser and Salamon's concerns follow from the general fact that accelerating rates of change increase intergenerational conflicts. Let me explain.

Here are three very long term historical trends:

1. Our total power and capacity has consistently increased. Long ago this enabled increasing population, and lately it also enables increasing individual income.
2. The rate of change in this capacity increase has also increased. This acceleration has been lumpy, concentrated in big transitions: from primates to humans to farmers to industry.
3. Our values, as expressed in words and deeds, have changed, and changed faster when capacity changed faster. Genes embodied many earlier changes, while culture embodies most today.

Increasing rates of change, together with constant or increasing lifespans, generically imply that individual lifetimes now see more change in capacity and in values. This creates more scope for conflict, wherein older generations dislike the values of younger more-powerful generations with whom their lives overlap.

As rates of change increase, these differences in capacity and values between overlapping generations increase. For example, Muehlhauser and Salamon fear that their lives might overlap with

> [descendants] superior to us in manufacturing, harvesting resources, scientific discovery, social charisma, and strategic action, among other capacities. We would not be in a position to negotiate with them, for [we] could not offer anything of value [they] could not produce more effectively themselves. … This brings us to the central feature of [descendant] risk: Unless a [descendant] is specifically programmed to preserve what [we] value, it may destroy those valued structures (including [us]) incidentally.

The quote actually used the words "humans", "machines" and "AI", and Muehlhauser and Salamon spend much of their chapter discussing the timing and likelihood of future AI. But those details are mostly irrelevant to the concerns expressed above. It doesn't matter much if our descendants are machines or biological meat, or if their increased capacities come from intelligence or raw physical power. What matters is that descendants could have more capacity and differing values.

Such intergenerational concerns are ancient, and in response parents have long sought to imprint their values onto their children, with modest success.

Muehlhauser and Salamon find this approach completely unsatisfactory. They even seem wary of descendants who are cell-by-cell emulations of prior human

brains, "brain-inspired AIs running on human-derived "spaghetti code", or 'opaque' AI designs …produced by evolutionary algorithms." Why? Because such descendants "may not have a clear 'slot' in which to specify desirable goals."

Instead Muehlhauser and Salamon prefer descendants that have "a transparent design with a clearly definable utility function," and they want the world to slow down its progress in making more capable descendants, so that they can first "solve the problem of how to build [descendants] with a stable, desirable utility function."

If "political totalitarians" are central powers trying to prevent unwanted political change using thorough and detailed control of social institutions, then "value totalitarians" are central powers trying to prevent unwanted value change using thorough and detailed control of everything value-related. And like political totalitarians willing to sacrifice economic growth to maintain political control, value totalitarians want us to sacrifice capacity growth until they can be assured of total value control.

While the basic problem of faster change increasing intergenerational conflict depends little on change being caused by AI, the feasibility of this value totalitarian solution does seem to require AI. In addition, it requires transparent-design AI to be an early and efficient form of AI. Furthermore, either all the teams designing AIs must agree to use good values, or the first successful team must use good values and then stop the progress of all other teams.

Personally, I'm skeptical that this approach is even feasible, and if feasible, I'm wary of the concentration of power required to even attempt it. Yes we teach values to kids, but we are also often revolted by extreme brainwashing scenarios, of kids so committed to certain teachings that they can no longer question them. And we are rightly wary of the global control required to prevent any team from creating descendants who lack officially approved values.

Even so, I must admit that value totalitarianism deserves to be among the range of responses considered to future intergenerational conflicts.