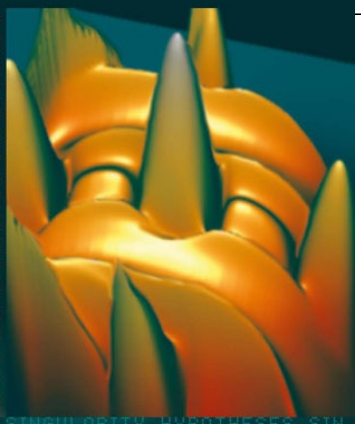
THE FRONTIERS COLLECTION



Amnon H. Eden · James H. Moor
Johnny H. Søraker · Eric Steinhart (Eds.)

SINGULARITY HYPOTHESES

A Scientific and Philosophical Assessment

Springer

Amnon H. Eden, James H. Moor, Johnny H. Søraker and Eric Steinhart (eds.), The Frontiers Collection, *Singularity Hypotheses*, 2012, A Scientific and Philosophical Assessment, DOI: 10.1007/978-3-642-32560-1, © Springer-Verlag Berlin Heidelberg 2012

The Frontiers Collection

The books in this collection are devoted to challenging and open problems at the forefront of modern science, including related philosophical debates. In contrast to typical research monographs, however they strive to present their topics in a manner accessible also to scientifically literate non-specialists wishing to gain insight into the deeper implications and fascinating questions involved. Taken as a whole, the series reflects the need for a fundamental and interdisciplinary approach to modern scienc Furthermore, it is intended to encourage active scientists in all areas to ponder over important and perhaps controversial issues beyond their own speciality. Extending from quantum physics and relativity to entropy, consciousness and complex systems—the Frontiers Collection will inspire readers to push back the frontiers of their own knowledge.

For a full list of published titles, please see back of book or springer.com/series/5342

For further volumes: http://www.springer.com/series/5342

Editors
Amnon H. Eden, James H. Moor, Johnny H. Søraker and Eric Steinhart

Singularity Hypotheses
A Scientific and Philosophical Assessment

Springer

Editors

Amnon H. Eden
School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK

James H. Moor
Dartmouth College, Hanover, USA

Johnny H. Søraker
Department of Philosophy, University of Twente, Enschede, The Netherlands

Eric Steinhart
Department of Philosophy, William Paterson University, Wayne, USA

ISSN 1612-3018

ISBN 978-3-642-32559-5 e-ISBN 978-3-642-32560-1

Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012946755

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Saul , With love ,

—Aba

Contents

Amnon H. Eden, James H. Moor, Johnny H. Søraker and Eric Steinhart (eds.), The Frontiers Collection, *Singularity Hypotheses*, 2012, A Scientific and Philosophical Assessment, DOI: 10.1007/978-3-642-32560-1_1,

1. Singularity Hypotheses: An Overview

Introduction to: Singularity Hypotheses: A Scientific and Philosophical Assessment

Amnon H. Eden[1], Eric Steinhart[2], David Pearce[3] and James H. Moor[4]

(1) School of Computer Science and, Electronic Engineering, University of Essex, Colchester, CO4 3SQ, UK
(2) Department of Philosophy, William Paterson University, 300 Pompton Road, Wayne, NJ 07470, USA
(3) knightsbridge Online, 7 Lower Rock Gardens, Brighton, USA
(4) Department of Philosophy, Dartmouth College, 6035 Thornton, Hanover, NH 03755, USA

Amnon H. Eden (Corresponding author)
Email: eden@essex.ac.uk

Eric Steinhart
Email: esteinhart1@nyc.rr.com

David Pearce
Email: dave@hedweb.com

James H. Moor
Email: james.h.moor@dartmouth.edu

Abstract
Bill Joy in a widely read but controversial article claimed that the most powerful 21st century technologies are threatening to make humans an endangered species. Indeed, a growing number of scientists, philosophers and forecasters insist that the accelerating progress in disruptive technologie such as artificial intelligence, robotics, genetic engineering, and nanotechnology may lead to what they refer to as the *technological singularity*: an event or phase that will radically change human civilization, and perhaps even human nature itself, before the middle of the 21st century.

Questions

Bill Joy in a widely read but controversial article claimed that the most powerful 21st century technologies are threatening to make humans an endangered species (Joy 2000). Indeed, a growing number of scientists, philosophers and forecasters insist that the accelerating progress in disruptive technologies such as artificial intelligence, robotics, genetic engineering, and nanotechnology may

lead to what they refer to as the *technological singularity*: an event or phase that will radically chang human civilization, and perhaps even human nature itself, before the middle of the 21st century (Paul and Cox 1996; Broderick 2001; Garreau 2005, Kurzweil 2005).

Singularity hypotheses refer to either one of two distinct and very different scenarios. The first (Vinge 1993; Bostrom to appear) postulates the emergence of artificial superintelligent agents—software-based synthetic minds—as the 'singular' outcome of accelerating progress in computing technology. This singularity results from an 'intelligence explosion' (Good 1965): a process in which software-based intelligent minds enter a 'runaway reaction' of self-improvement cycles, with each new and more intelligent generation appearing faster than its predecessor. Part I of this volume is dedicated to essays which argue that progress in artificial intelligence and machine learning may indeed increase machine intelligence beyond that of any human being. As Alan Turing (1951) observed, "at some stage therefore we should have to expect the machines to take control, in the way that is mentioned in Samuel Butler's 'Erewhon' ": the consequences of such greater-than-human intelligence will be profound, and conceivably dire for humanity as we know it. Essays in Part II of this volume are concerned with this scenario.

A radically different scenario is explored by transhumanists who expect progress in enhancement technologies, most notably the amplification of human cognitive capabilities, to lead to the emergenc of a posthuman race. Posthumans will overcome all existing human limitations, both physical and mental, and conquer aging, death and disease (Kurzweil 2005). The nature of such a singularity, a 'biointelligence explosion', is analyzed in essays in Part III of this volume. Some authors (Pearce, th volume) argue that transhumans and posthumans will retain a fundamental biological core. Other authors argue that fully functioning, autonomous whole-brain emulations or 'uploads' (Chalmers 2010; Koene this volume; Brey this volume) may soon be constructed by 'reverse-engineering' the brain of any human. If fully functional or even conscious, uploads may usher in an era where the notion of personhood needs to be radically revised (Hanson 1994).

Advocates of the technological singularity have developed a powerful inductive *Argument from Acceleration* in favour of their hypothesis. The argument is based on the extrapolation of *trend curve* in computing technology and econometrics (Moore 1965; Moravec 1988, Chap. 2; Moravec 2000, Chap. 3; Kurzweil 2005, Chaps. 1 and 2). In essence, the argument runs like this: (1) The study of the history of technology reveals that technological progress has long been accelerating. (2) There are good reasons to think that this acceleration will continue for at least several more decades. (3) If it does continue, our technological achievements will become so great that our bodies, minds, societies and economies will be radically transformed. (4) Therefore, it is likely that this disruptive transformation will occur. Kurzweil (2005, p. 136) sets the date mid-century, around the year 2045. The change will be so revolutionary that it will constitute a "rupture in the fabric of human history" (Kurzweil 2005, p. 9).

Critics of the technological singularity dismiss these claims as speculative and empirically unsound, if not pseudo-scientific (Horgan 2008). Some attacks focus on the premises of the Argumer from Acceleration (Plebe and Perconti this volume), mostly (2). For example, Modis (2003; this volume) claims that after periods of change that appear to be accelerating, technological progress always levels off. Other futurists have long argued that we are heading instead towards a global economic and ecological collapse. This negative scenario was famously developed using computer modelling of the future in *The Limits to Growth* (Meadows et al. 1972, 2004).

Mocked as the "rapture of the nerds", many critics take (3) to be yet another apocalyptic fantasy, technocratic variation on the usual theme of doom-and-gloom fuelled by mysticism, science fiction and even greed. Some conclude that the singularity is a religious notion, not a scientific one (Horgan 2008; Proudfoot this volume; Bringsjord et al. this volume). Other critics (Chaisson this volume)

accept acceleration as an underlying law of nature but claim that, in perspective, the significance of the claimed changes is overblown. That is, what is commonly described as the technological singularity may well materialize, with profound consequences for the human race. But on a cosmic scale, such a mid-century transition is no more significant then whatever may follow.

Existential risk or cultist fantasy? Are any of the accounts of the technological singularity credible? In other words, is the technological singularity an open problem in science?

We believe that before any interpretation of the singularity hypothesis can be taken on board by the scientific community, rigorous tools of scientific enquiry must be employed to reformulate it as a coherent and falsifiable conjecture. To this end, we challenged economists, computer scientists, biologists, mathematicians, philosophers and futurists to articulate their concepts of the singularity. The questions we posed were as follows:

1. What is the [technological] singularity hypothesis? What exactly is being claimed?

2. What is the empirical content of this conjecture? Can it be refuted or corroborated, and if so, how?

3. What exactly is the nature of a singularity: Is it a discontinuity on a par with phase transition or a process on a par with Toffler's 'wave'? Is the term singularity appropriate?

4. What evidence, taken for example from the history of technology and economic theories, suggest the advent of some form of singularity by 2050?

5. What, if anything, can be said to be accelerating? What evidence can reliably be said to support its existence? Which metrics support the idea that 'progress' is indeed accelerating?

6. What are the most likely milestones ('major paradigm shifts') in the countdown to a singularity?

7. Is the so-called Moore's Law on par with the laws of thermodynamics? How about the Law of Accelerating Returns? What exactly is the nature of the change they purport to measure?

8. What are the necessary and sufficient conditions for an intelligence explosion (a *runaway effect*)? What is the actual likelihood of such an event?

9. What evidence support the claim that machine intelligence has been rising? Can this evidence be extrapolated reliably?

10. What are the necessary and sufficient conditions for machine intelligence to be considered to be on a par with that of humans? What would it take for the "general educated opinion [to] have altered so much that one will be able to speak of machines thinking without expecting to be contradicted" (Turing 1950, p. 442)?

11. What does it mean to claim that biological evolution will be replaced by technological evolution? What exactly can be the expected effects of augmentation and enhancement, in particular over our cognitive abilities? To which extent can we expect our transhuman and posthuman descendants to be different from us?

12. What evidence support the claim that humankind's intelligence quotient has been rising ("Flynn effect")? How this evidence relate to a more general claim about a rise in the 'intelligence' of carbon-based life? Can this evidence be extrapolated reliably?

13. What are the necessary and sufficient conditions for a functioning whole brain emulation (WBE) of a human? At which level exactly must the brain be emulated? What will be the conscious experience of a WBE? To which extent can they be said to be human?

14. What may be the consequences of a singularity? What may be its effect on society, e.g. in ethics, politics, economics, warfare, medicine, culture, arts, the humanities, and religion?

15. Is it meaningful to refer to multiple singularities? If so, what can be learned from past such events? Is it meaningful to claim a narrow interpretation of singularity in some specific domain of activity, e.g. a singularity in chess playing, in face recognition, in car driving, etc.?

This volume contains the contributions received in response to this challenge.

Towards a Definition

Accounts of a technological singularity—henceforth the singularity—appear to disagree on its causes and possible consequences, on timescale, and even on its nature: the emergence of machine intelligence or of posthumans? An event or a period? Is the technological singularity unique or have there been others? The absence of a consensus on basic questions casts doubt whether the notion of *singularity* is at all coherent.

The term in its contemporary sense traces back to von Neumann, who is quoted as saying that "th ever-accelerating progress of technology and changes in the mode of human life… gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue" (in Ulam 1958). Indeed, the twin notions of *acceleration* and *discontinuity* are common to all accounts of the technological singularity, as distinguished from a space–time singularity and a singularity in a mathematical function.

Acceleration refers to a rate of growth in some quantity such as computations per second per fixe dollar (Kurzweil 2005), economic measures of *growth rate* (Hanson 1994; Miller this volume) or tota output of goods and services (Toffler 1970), and *energy rate density* (Chaisson this volume). Others describe quantitative measures of physical, biological, social, cultural, and technological processes o evolution: milestones or 'paradigm shifts' whose timing demonstrates an accelerating pace of change For example, Sagan's Cosmic Calendar (1977, Chap. 1) names milestones in biological evolution suc as the emergence of eukaryotes, vertebrates, amphibians, mammals, primates, *hominidae*, and *Homo sapiens*, which show an accelerating trend. Following Good (1965) and Bostrom (to appear), Muehlhauser and Salamon (this volume), Arel (this volume), and Schmidhuber (this volume) describ developments in machine learning which seek to demonstrate that progressively more 'intelligent' problems have been solved during the past few decades, and how such technologies may further improve, possibly even in a recursive process of self-modification. Some authors attempt to show tha many of the above accounts of acceleration are in fact manifestations of an underlying law of nature (Adams 1904; Kurzweil 2005, Chaisson this volume): quantitatively or qualitatively measured, acceleration is commonly visualized as an upwards-curved mathematical graph which, if projected into the future, is said to be leading to a *discontinuity*.

Described either as an event that may take a few hours (e.g., a 'hard takeoff', Loosemore and Goertzel, this volume) or a period of years (e.g., Toffler 1970), the technological singularity is taken to mark a *discontinuity* or a turning-point in human history. The choice of word 'singularity' appears to be motivated less by the eponymous mathematical concept (Hirshfeld 2011) and more by the ontological and epistemological discontinuities idiosyncratic to black holes. Seen as a central metaphor, a gravitational singularity is a (theoretical) point at the centre of black holes at which quantities that are otherwise meaningful (e.g., *density* and *spacetime curvature*) become infinite, or rather meaningless. The discontinuity expressed by the black hole metaphor is thus used to convey how the quantitative measure of *intelligence*, at least as it is measured by traditional IQ tests (such as Wechsler and Stanford-Binet), may become a meaningless notion for capturing the intellectual capabilities of superintelligent minds. Alternatively, we may say a graph measuring average intelligence beyond the singularity in terms of IQ score may display some form of radical discontinuity if superintelligence emerges. Furthermore, singularitarians note that gravitational singularities are said to be surrounded by an *event horizon*: a boundary in spacetime beyond which events cannot be observed from outside, and a horizon beyond which gravitational pull becomes so strong that nothing can escape, even light (hence "black")—a point of no return. Kurzweil (2005) and others (e.g., Pearce this volume) contend that, since the minds of superintelligent intellects may be difficult or impossible for humans to comprehend (Fox and Yampolskiy this volume), a technological singularity marks an epistemological barrier beyond which events cannot be predicted or understood—an 'event horizon' in human affairs. The gravitational singularity metaphor thus reinforces the view that the change will be radical and that its outcome cannot be foreseen.

The combination of acceleration and discontinuity is at once common and unique to the singularity literature in general and to the essays in this volume in particular. We shall therefore proceed on the premise that acceleration and discontinuity jointly offer necessary and sufficient conditions for us to take a manuscript to be concerned with a hypothesis of a technological singularity.

Historical Background

Many philosophers have portrayed the cosmic process as an ascending curve of positivity (Lovejoy 1936, Chap. 9). Over time, the quantities of intelligence, power or value are always increasing. These progressive philosophies have sometimes been religious and sometimes secular. Secular versions of progress have sometimes been political and sometimes technological. Technological versions have sometimes invoked broad technical progress and have sometimes focused on more specific outcomes such as the possible recursive self-improvement of artificial intelligence.

For some philosophers of progress, the rate of increase remains relatively constant; for others, the rate of increase is also increasing—progress accelerates. Within such philosophies, the *singularity* is often the point at which positivity becomes maximal. It may be an ideal limit point (an *omega point*) either at infinity or at the vertical asymptote of an accelerating trajectory. Or sometimes, the singularity is the critical point at which the slope of an accelerating curve passes beyond unity.

Although thought about the singularity may appear to be very new, in fact such ideas have a long philosophical history. To help increase awareness of the deep roots of singularitarian thought within traditional philosophy, it may be useful to look at some of its historical antecedents.

Perhaps the earliest articulation of the idea that history is making progress toward some omega point of superhuman intelligence is found in *The Phenomenology of Spirit*, written by Hegel (1807). Hegel describes the ascent of human culture to an ideal limit point of absolute knowing. Of course, Hegel's thought is not technological. Yet it is probably the first presentation, however abstract, of singularitarian ideas. For the modern Hegelian, the singularity looks much like the final self-

realization of Spirit in absolute knowing (Zimmerman 2008).

Around 1870, the British writer Samuel Butler used Darwinian ideas to develop a theory of the evolution of technology. In his essay "Darwin among the Machines" and in his utopian novel *Erewhon: Or, Over the Range* (Butler 1872), Butler argues that machines would soon evolve into artificial life-forms far superior to human beings. Threatened by superhuman technology, the Erewhonians are notable for rejecting all advanced technology. Also writing in the late 1800s, the American philosopher Charles Sanders Peirce developed an evolutionary cosmology (see Hausman 1993). Peirce portrays the universe as evolving from an initial chaos to a final singularity of pure mind. Its evolution is accelerating as this tendency to regularity acts upon itself. Although Pierce's notion of progress was not based on technology, his work is probably the earliest to discuss the notio of accelerating progress itself. Of course, Peirce was also a first-rate logician; and as such, he was among the first to believe that minds were computational machines.

Around 1900, the American writer Henry Adams [1] was probably the first writer to describe a technological singularity. Adams was almost certainly the first person to write about history as a self accelerating technological process. His essay "The Law of Acceleration" (Adams 1904) may well be the first work to propose an actual formula for the acceleration of technological change. Adams suggests measuring technological progress by the amount of coal consumed by society. His law of acceleration prefigures Kurzweil's law of accelerating returns. His later essay "The Rule of Phase" (Adams 1909) portrays history as accelerating through several epochs—including the Instinctual, Religious, Mechanical, Electrical, and Ethereal Phases. This essay contains what is probably the first illustration of history as a curve approaching a vertical asymptote. Adams provides a mathematical formula for computing the duration of each technological phase, and the amount of energy that will consumed during that phase. His epochs prefigure Kurzweil's evolutionary epochs. Adams uses his formulae to argue that the singularity will be reached by about the year 2025, a forecast remarkably close to modern singularitarians.

Much writing on the singularity owes a great debt to Teilhard de Chardin (1955; see Steinhart 2008). Teilhard is among the first writers seriously to explore the future of human evolution. He advocates both biological enhancement technologies and artificial intelligence. He discusses the emergence of a global computation-communication system (and is said by some to have been the firs to have envisioned the Internet). He proposes the development of a global society and describes the acceleration of progress towards a technological singularity (which he termed "the critical point"). H discusses the spread of human intelligence into the universe and its amplification into a cosmic-intelligence. Much of the more religiously-expressed thought of Kurzweil (e.g. his definition of "Goo as the omega point of evolution) ultimately comes from Teilhard.

Many of the ideas presented in recent literature on the singularity are foreshadowed in a prescien essay by George Harry Stine. Stine was a rocket engineer and part-time science fiction writer. His essay "Science Fiction is too Conservative" was published in May 1961 in *Analog*. *Analog* was a widely read science-fiction magazine. Like Adams, Stine uses trend curves to argue that a momentou and disruptive event is going to happen in the early 21st Century.

In 1970, Alvin and Heidi Toffler observed both acceleration and discontinuity in their influential work, *Future Shock*. About acceleration, the Tofflers observed that "the total output of goods and services in advanced societies doubles every 15 years, and that the doubling times are shrinking" (Toffler 1970, p. 25). They demonstrate accelerating change in every aspect of modern life: in transportation, size of population centres, family structure, diversity of lifestyles, etc., and most importantly, in the transition from factories as 'means of production' to knowledge as the most fundamental source of wealth (Toffler 1980). The Tofflers conclude that the transition to knowledge-based society "is, in all likelihood, bigger, deeper, and more important than the industrial revolution.

… Nothing less than the second great divide in human history, the shift from barbarism to civilization" (Toffler 1970, p. 11).

During the 1980s, unprecedented advances in computing technology led to renewed interest in the notion that technology is progressing towards some kind of tipping-point or discontinuity. Moravec's *Mind Children* (1988) revived research into the nature of technological acceleration. Many more books followed, all arguing for extraordinary future developments in robotics, artificial intelligence, nanotechnology, and biotechnology. Kurzweil (1999) developed his law of accelerating returns in *The Age of Spiritual Machines*. Broderick (2001) brought these ideas together to argue for a future climax of technological progress that he termed *the spike*. All these ideas were brought into public consciousness with the publication of Kurzweil's (2005) *The Singularity is Near* and its accompanying movie. As the best-known defence of the singularity, Kurzweil's work inspired dozens of responses. One major assessment of singularitarian ideas was delivered by *Special Report: The Singularity* in *IEEE Spectrum* (June, 2008). More recently, notable work on the singularity has been done by the philosopher David Chalmers (2010) and the discussion of the singularity it inspired (*The Journal of Consciousness Studies* 19, pp. 1–2). The rapid growth in singularity research seems set to continue and perhaps accelerate.

Essays in this Volume

The essays developed by our authors divide naturally into several groups. Essays in Part I hold that a singularity of machine superintelligence is probable. Luke Muehlhauser and Anna Salamon of the Singularity Institute of Artificial Intelligence argue that an intelligence explosion is likely and examine some of its consequences. They make recommendations designed to ensure that the emerging superintelligence will be beneficial, rather than detrimental, to humanity. Itamer Arel, a computer scientist, argues that artificial general intelligence may become an extremely powerful and disruptive force. He describes how humans might shape the emergence of superhuman intellects so that our relations with such intellects are more cooperative than competitive. Juergen Schmidhuber, also a computer scientist, presents substantial evidence that improvements in artificial intelligence are rapidly progressing towards human levels. Schmidhuber is optimistic that, if future trends continue, we will face an intelligence explosion within the next few decades. The last essay in this part is by Richard Loosemore and Ben Goertzel who examine various objections to an intelligence explosion and conclude that they are not persuasive.

Essays in Part II are concerned with the values of agents that may result from a singularity of artificial intellects. Luke Muehlhauser and Louie Helm ask what it would mean for artificial intellect to be *friendly* to humans, conclude that human values are complex and difficult to specify, and discuss techniques we might use to ensure the friendliness of artificial superintelligent agents. Joshua Fox and Roman Yampolskiy consider the psychologies of artificial intellects. They argue that human-like mentalities occupy only a very small part of the space of possible minds. If Fox and Yampolskiy are right, then it is likely that such minds, especially if superintelligent, will scarcely be recognizable to us at all. The values and goals of such minds will be alien, and perhaps incomprehensible in human terms. This strangeness creates challenges, some of which are discussed in James Miller's essay. Miller examines the economic issues associated with a singularity of artificial superintelligence. He shows that although the singularity of artificial superintelligence may be brought about by economic competition, one paradoxical consequence might be the destruction of the value of money. More worryingly, Miller suggests that a business that may be capable of creating an artificial superintelligence would face a unique set of economic incentives likely to push it deliberately to make it *unfriendly*. To counter such worries, Steve Omohundro examines how market forces may affect

their behaviour. Omohundro proposes a variety of strategies to ensure that any artificial intellects wi have human-friendly values and goals. Eliezer Yudkowsky concludes this part by considering the ways that artificial superintelligent intellects may radically differ from humans and the urgent need for us to take those differences into account.

Whereas essays in Parts I and II are concerned with the intelligence explosion scenario—a singularity deriving from the evolution of intelligence in silicon, the essays in Part III are concerned with the evolution that humans may undergo via enhancement, amplification, and modification, and with the scenario in which a race of superintelligent posthumans emerges. David Pearce conceives of humans as 'recursively self-improving organic robots' poised to re-engineer their own genetic code and bootstrap their way to full-spectrum superintelligence. Hypersocial and supersentient, the successors of archaic humanity may phase out the biology of suffering throughout the living world. Randal Koene examines how the principles of evolution apply to brain emulations. He argues that intelligence entails autonomy, so that future 'substrate-independent minds' (SIMs), may hold values that humans find alien. Koene nonetheless hopes that, since SIMs will originate from our own brains human values play significant roles in superintelligent, 'disembodied' minds. Dennis Bray examines the biochemical mechanisms of the brain. He concludes that building fully functional emulations by reverse-engineering human brains may entail much more than modelling neurons and synapses. However, there are other ways to gain inspiration from the evolution of biological intelligence. We may be able to harness brain physiology and natural selection to evolve new types of intelligence, an perhaps superhuman intelligence. David Roden worries that the biological moral heritage of humanit may disappear entirely after the emergence of superintelligent intellects, whether artificial or of biological origin. Such agents may emerge with utterly novel features and behaviour that cannot be predicted from their evolutionary histories.

The essays in Part IV of the volume are skeptical about the singularity, each focusing on a particular aspect such as the intelligence explosion or the prospects of acceleration continuing over the next few decades. A report developed by the American Association for Artificial Intelligence considers the future development of artificial intelligence (AI). While optimistic about specific advances, the report is highly skeptical about grand predictions of an intelligence explosion, of a 'coming singularity', and about any loss of human control. Alessio Plebe and Pietro Perconti argue that the trends analysis as singularitarians present it is faulty: far from rising, the pace of change is n accelerating but in fact slowing down, and even starting to decline. Futurist Theodore Modis is deepl skeptical about any type of singularity. He focuses his skepticism on Kurzweil's work, arguing that analysis of past trends does not support long-term future acceleration. For Modis, technological change takes the form of S-curves (logistic functions), which means that its trajectory is consistent with exponential acceleration for only a very short time. Modis expects computations and related technologies to slow down and level off. While technological advances will continue to be disruptive there will be no singularity. Other authors go further and argue that most literature on the singularity is not genuinely scientific but theological. Focusing on Kurzweil's work, Diane Proudfoot's essay develops the notion that singularitarianism is a kind of millenarian ideology (Bozeman 1997; Geraci 2010; Steinhart 2012) or "the religion of technology" (Noble 1999). Selmer Bringsjord, Alexander Bringsjord, and Paul Bello compare belief in the singularity to fideism in traditional Christianity, which denies the relevance of evidence or reason.

The last essay in Part IV offers an ambitious theory of acceleration that attempts to unify cosmic evolution with biological, cultural and technological evolution. Eric Chaisson argues that complexity can be shown consistently to increase from the Big Bang to the present, and that the same forces that drive the rise of complexity in Nature generally also underlie technological progress. To support this sweeping argument, Chaisson defines the physical quantity of *energy density rate* and shows how it

unifies the view of an accelerating grow along physical, biological, cultural, and technological evolution. But while Chaisson accepts the first element of the technological singularity, *acceleration* he rejects the second, *discontinuity*—hence the singularity: "there is no reason to claim that the next evolutionary leap forward beyond sentient beings and their amazing gadgets will be any more important than the past emergence of increasingly intricate complex systems." Chaisson reminds us that our little planet is not the only place in the universe where evolution is happening. Our machines may achieve superhuman intelligence. But perhaps a technological singularity will happen first elsewhere in the cosmos. Maybe it has already done so.

Conclusions

History shows time and again that the predictions made by futurists (and economists, sociologists, politicians, etc.) have been confounded by the behaviour of self-reflexive agents. Some forecasts are self-fulfilling, others self-stultifying. Where, if at all, do predictions of a technological singularity fi into this typology? How are the lay public/political elites likely to respond if singularitarian ideas ga widespread currency? Will the 21st century mark the end of the human era? And if so, *will biologica humanity's successors be our descendants*? It is our hope and belief that this volume will help to mo these questions beyond the sometimes wild speculations of the blogosphere and promote the growth singularity studies as a rigorous scholarly discipline.

References

Adams, H. (1909). The rule of phase applied to history. In H. Adams & B. Adams (Eds.) (1920) *The degradation of the democratic dogma*. (pp. 267–311). New York: Macmillan.

Adams, H. (1904). A law of acceleration. In H. Adams (1919) *The education of Henry Adams*. New York: Houghton Mifflin, Chap. 34.

Bostrom, N. to appear. "Intelligence explosion".

Bozeman. J. (1997). Technological millenarianism in the United States. In T. Robbins & S. Palmer (Eds.) (1997) *Millennium, messiahs, and mayhem: contemporary apocalyptic movements* (pp. 139–158). New York: Routledge.

Butler, S. (1872/1981). *Erewhon: or, over the range*. In H. P. Breuer & D. F. Howard. Newark: University of Delaware Press.

Broderick, D. (2001). The spike: how our lives are being transformed by rapidly advancing technologies. New York: Tom Doherty Associates.

Chalmers, D. (2010). The singularity: a philosophical analysis. *Journal of Consciousness Studies 17*, 7–65.

Garreau, J. (2005). *Radical evolution*. New York: Doubleday.

Geraci, R. (2010). Apocalypic AI: visions of heaven in robotics, artificial intelligence, and virtual reality. New York: Oxford University Press.

Good, I. (1965). Speculations concerning the first ultraintelligent machine. In Alt, F., Rubinoff, M. (Eds.) *Advances in Computers* Vol. 6. New York: Academic Press.

Hanson, R. (1994). If uploads come first: crack of a future dawn. *Extropy 6* (1), 10–15.

Hausman, C. (1993). *Charles S. Peirce's evolutionary philosophy*. New York: Cambridge University Press.

Hirshfeld, Y. (2011). A note on mathematical singularity and technological singularity. *The singularity hypothesis*, blog entry, 5 Fe Available http://singularityhypothesis.blogspot.co.uk/2011/02/note-on-mathematical-singularity-and.html.

Horgan, J. (2008). The consciousness conundrum. *IEEE Spectrum 45*(6), 36–41.

Hegel, G. W. F. (1807/1977). *Phenomenology of spirit.* Trans. A. V. Miller. New York: Oxford University Press.

Joy, B. (2000). Why the future doesn't need us. http://www.wired.com/wired/archive/8.04/joy.html.

Kurzweil, R. (1999). *The age of spiritual machines.* New York: Penguin.

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

Lovejoy, A. (1936). *The great chain of being.* Cambridge: Harvard University Press.

Meadows, D. H., Meadows, D. L, Randers, J., & Behrens, W. (1972). *The limits to growth: a report for the club of Rome project on the predicament of mankind.* New York: Universe Books.

Meadows, D. H., Randers, J., & Meadows, D. L (2004). *The limits to growth: the thirty year update.* White River Junction, VT: Chelsea Green Books.

Modis, T. (2003). The limits of complexity and change. *The futurist* (May–June), 26–32.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics 38*(8), 114–117.

Moravec, H. (1988). *Mind children: the future of robot and human intelligence.* Cambridge: Harvard University Press.

Moravec, H. (2000). *Robot: mere machine to transcendent mind*. New York: Oxford University Press.

Noble, D. F. (1999). The religion of technology: the divinity of man and the spirit of invention. New York: Penguin.

Paul, G. S., E. D. Cox (1996). *Beyond humanity: cyberevolution and future minds.* Rockland, MA: Charles River Media.

Sagan, C. (1977). *The dragons of Eden.* New York: Random House.

Steinhart, E. (2008). Teilhard de Chardin and transhumanism. *Journal of Evolution and Technology 20*, 1–22. Online at < jetpress.org/v20/steinhart.htm >.

Steinhart, E. (2012). Digital theology: is the resurrection virtual? In M. Luck (Ed.) (2012) *A philosophical exploration of new and alternative religious movements.* Farnham, UK: Ashgate.

Stine, G. H. (1961). Science fiction is too conservative. *Analog Science Fact and Fiction LXVII* (3), 83–99.

Teilhard de Chardin, P. (1955/2002). *The phenomenon of man*. Transactions B. Wall. New York: Harper Collins. Originally written 1938–1940.

Toffler, A. (1970). *Future shock.* New York: Random House.

Toffler, A. (1980). *The third wave*. New York: Bantam.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind 59* (236), 433–460.

Turing, A. M. (1951). Intelligent machinery, a heretical theory. *The 51 Society*. BBC programme.

Ulam, S. (1958) Tribute to John von Neumann. *Bulletin of the American mathematical society 64* (3.2), 1–49.

Vinge, V. (1993). The coming technological singularity: how to survive in the post-human era. In *Proc. Vision 21: interdisciplinary science and engineering in the era of cyberspace*, (pp. 11–22). NASA: Lewis Research Center.

Zimmerman, M. (2008). The singularity: a crucial phase in divine self-actualization? *Cosmos and history: The Journal of Natural a Social Philosophy 4*(1–2), 347–370.

Footnotes

descendant of President John Quincy Adams.

Part 1

A Singularity of Artificial Superintelligence

Amnon H. Eden, James H. Moor, Johnny H. Søraker and Eric Steinhart (eds.), The Frontiers Collection, *Singularity Hypotheses*, 2012, A Scientific and Philosophical Assessment, DOI: 10.1007/978-3-642-32560-1, © Springer-Verlag Berlin Heidelberg 2012

Amnon H. Eden, James H. Moor, Johnny H. Søraker and Eric Steinhart (eds.), The Frontiers Collection, *Singularity Hypotheses*, 2012, A Scientific and Philosophical Assessment, DOI: 10.1007/978-3-642-32560-1_2, © Springer-Verlag Berlin Heidelberg 2012

2. Intelligence Explosion: Evidence and Import

Luke Muehlhauser[1] and Anna Salamon[1]

(1) Machine Intelligence Research Institute, Berkeley, CA 94705, USA

Luke Muehlhauser
Email: luke@intelligence.org

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, b
only briefly
Vernor Ving

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 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 wi 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 i 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 trajectorie 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 ar 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 preference 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 optimizatio 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).

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) bu merely a *narrow* AI, because it can only solve a narrow set of problems. Drop Watson in a pond or as it to do original science, and it would be helpless even if given a month's warning to prepare. Imagin