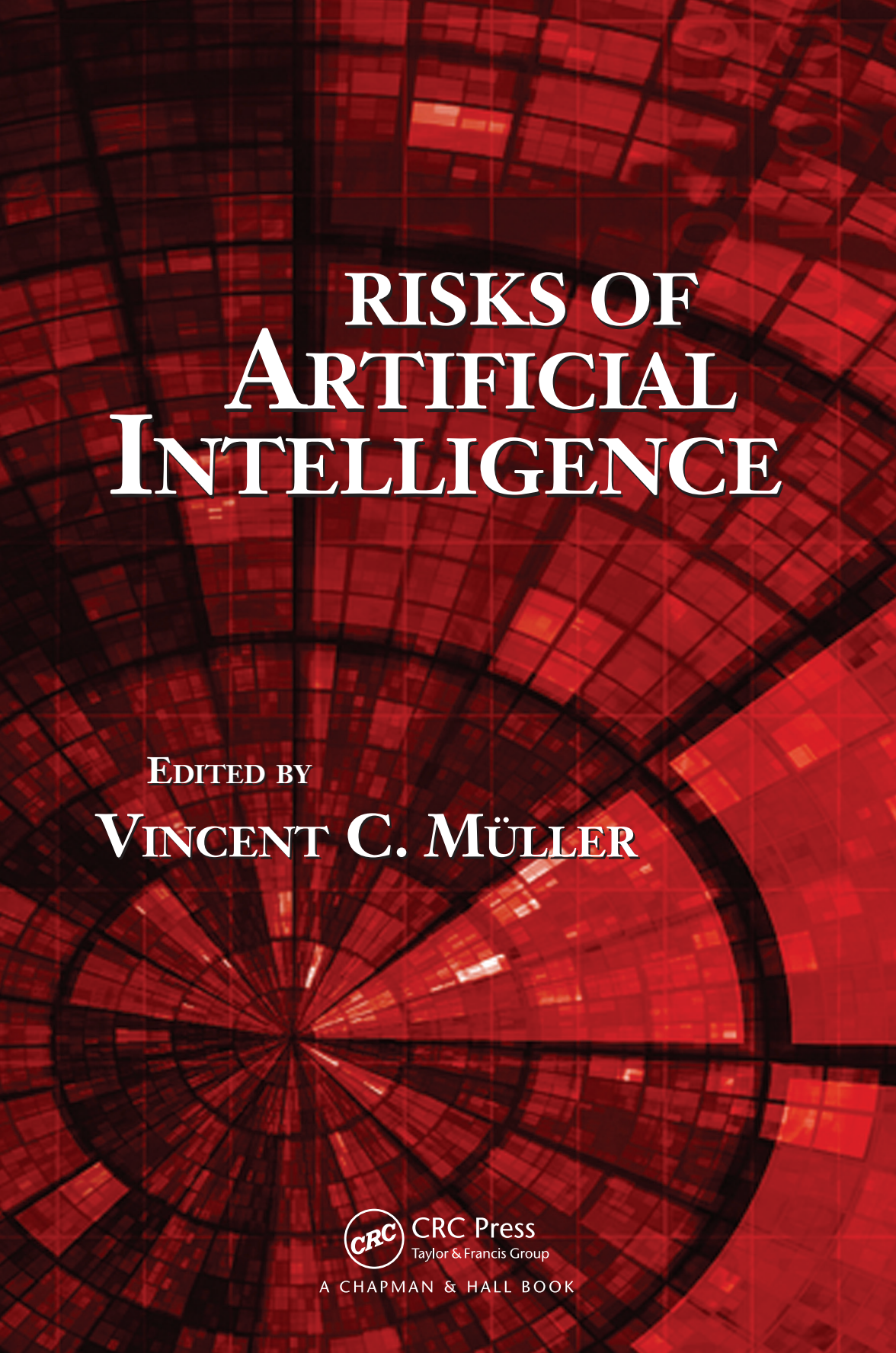
RISKS OF
ARTIFICIAL
INTELLIGENCE
EDITED BY
VINCENT C. MÜLLER
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

RISKS OF ARTIFICIAL INTELLIGENCE

RISKS OF ARTIFICIAL INTELLIGENCE

EDITED BY

VINCENT C. MÜLLER

UNIVERSITY OF OXFORD, UK

AND

AMERICAN COLLEGE OF THESSALONIKI/ANATOLIA COLLEGE, GREECE

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20151012

International Standard Book Number-13: 978-1-4987-3483-7 (eBook - PDF)

Contents

Editor

Vincent C. Müller's research focuses on the nature and future of computational systems, particularly on the prospects and dangers of artificial intelligence. He is the president of the European Association for Cognitive Systems and was the coordinator of the European Network for Cognitive Systems, Robotics and Interaction, which has nearly 1000 members and is funded by the European Commission through two FP7 projects worth €3.9 million over 2009–2014 (www.eucognition.org). He organizes a conference series, Theory and Philosophy of AI (www.pt-ai.org) and is the principal investigator of a European Commission–funded research project Digital DIY. He is currently working as professor of philosophy, Division of Humanities & Social Sciences, Anatolia College/ACT, Pylaia-Thessaloniki, Greece.

Müller has published a number of articles in leading journals on the philosophy of computing, the philosophy of artificial intelligence and cognitive science, and the philosophy of language, applied ethics, and related areas. He has edited 10 volumes, mostly on the theory of cognitive systems and artificial intelligence, and is preparing a monograph on the fundamental problems of artificial intelligence. Müller studied philosophy with cognitive science, linguistics, and history at the universities of Marburg, Hamburg, London, and Oxford. He was Stanley J. Seeger Visiting Fellow at Princeton University and James Martin Research Fellow at the University of Oxford.

ORCID: 0000-0002-4144-4957

Contributors

Stuart Armstrong
The Future of Humanity Institute
University of Oxford
Oxford, UK

J. Mark Bishop
Cognitive Computing at Goldsmiths
and
The Goldsmiths Centre for Intelligent Data Analytics (TCIDA)
University of London
London, UK

Miles Brundage
Consortium for Science, Policy, and Outcomes
Arizona State University
Tempe, Arizona

Daniel Dewey
The Future of Humanity Institute
Oxford Martin School
University of Oxford
Oxford, UK

Ben Goertzel
Novamente LLC
Rockville, Maryland

Ted Goertzel
Rutgers University
Camden, New Jersey

András Kornai
Computer and Automation Research Institute
Hungarian Academy of Sciences
Budapest, Hungary

and

Department of Computer Science
Boston University
Boston, Massachusetts

Seán S. ÓhÉigeartaigh
The Future of Humanity Institute
University of Oxford
Oxford, UK

Steve Omohundro
Self-Aware Systems
Palo Alto, California

Alexey Potapov
AIDEUS
St. Petersburg, Russia
and
National Research University
of Information Technology,
Mechanics and Optics
St. Petersburg, Russia

Sergey Rodionov
AIDEUS
St. Petersburg, Russia

and

Aix Marseille Université, CNRS
LAM (Laboratoire d'Astrophysique
de Marseille)
Marseille, France

CHAPTER 1

Editorial

Risks of Artificial Intelligence

Vincent C. Müller

CONTENTS

ABSTRACT If the intelligence of artificial systems were to surpass that of humans significantly, humanity would face a significant risk. The time has come to consider these issues, and this consideration must include progress in artificial intelligence (AI) as much as insights from the theory of AI. The chapters in this volume try to make cautious headway in setting the problem, evaluating predictions on the future of AI, proposing ways to ensure that AI systems will be beneficial to humans, and critically evaluating such proposals.

1.1 INTRODUCTION: RISK OF ARTIFICIAL INTELLIGENCE

This is the first volume of papers ever dedicated to the risks of AI and it originates from the first conference on the risks of AI (AGI Impacts, Oxford, December 2012). Following that conference, was published a volume of selected papers in the *Journal of Experimental and Theoretical Artificial Intelligence* (*JETAI*) (see Müller, 2014). Our volume generated significant interest: there were approximately 20,000 paper downloads from the *JETAI* site alone in the first year, and three of the top five downloaded

papers from *JETAI* are now from our volume. As a result, the publishers suggested turning the journal volume into a book after adding some recent material, so this is what you have in front of you.

The notion that AI might generate an existential threat to humanity has gained currency since we published the journal volume: Nick Bostrom's book *Superintelligence: Paths, Dangers, Strategies* has come out (Bostrom, 2014); well-known public intellectuals such as Stephen Hawking and Stuart Russell have published warning notes in the general press (Hawking et al., 2014); and a host of media publications has followed. The idea of existential risk, to use a Hollywood cliché, is *the machines will take over and kill us all*—this fear obviously strikes a cord. The spread of this fear has generated significant concern among academics in AI. One indication is that the current and past presidents of the Association for the Advancement of Artificial Intelligence (that most significant academic AI association) wrote a short statement to the effect that "AI doomsday scenarios belong more in the realm of science fiction than science fact", although they also "urge our colleagues in industry and academia to join us in identifying and studying these risks" (Ditterich and Horowitz, 2015). Recently, some efforts have been made to outline the research agenda for AI that is beneficial for humanity, for example, in the new "Future of Life Institute" (Russell et al., 2015) and the renamed "Machine Intelligence Research Institute (MIRI)" (Soares and Fallenstein, 2014).

Although the traditional concerns in the philosophy and theory of AI have focused on the prospects of AI, its relation to cognitive science, and its fundamental problems (see conference series at www.pt-ai.org), we now see an increasing focus on matters of risk and ethics. But what is the general idea of this shift?

1.2 RISKS OF AI

The notion of an agent with general intelligence ability is the original driving vision of AI research (see McCarthy et al., 1955) and dominates much of its public image—although nearly all actual current work in AI is on specialized technology, far removed from such a general ability and often without use of the term "artificial intelligence."

The move from AI to risk is relatively easy: there is no reason to think that the level of human intelligence is anything special in the space of possibilities—it is easy to imagine natural or artificial intelligence agents that are vastly superior to us. There also seem to be reasons to think that the development of AI is accelerating, together with related technologies,

and that the invention of intelligent machines itself would further accelerate this development, thus constituting an "argument from acceleration" for the hypothesis that some disruptive transformation will occur. If one thinks of intelligence as a quantifiable unit, then this acceleration will continue and move past the (small) space that marks the intelligence of humans. Therefore, we will reach "superintelligence," which Bostrom tentatively defines as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom, 2014, p. 22). In a classic passage, Good has speculated that "the first ultraintelligent machine is the *last* invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control" (Good, 1965, section 2). Therefore, there is the risk that "the machines take over" and this loss of control is a significant risk, perhaps an existential risk for humanity [for a survey, see Sotala and Yampolskiy (2013)].

The discussion of risk is *not* dependent on the view that AI is now on a successful path toward superintelligence—though it gains urgency if such "success" is a nonnegligible possibility in the coming decades. It also gains urgency if the stakes are set high, even up to human extinction. If the stakes are so high as to include extinction of humankind, even a fairly small possibility of a disastrous outcome (say, 3%) is entirely sufficient to motivate the research. Consider that if there were a 3% possibility that a plane you are about to board will crash: that would be sufficient motivation for getting off. The utility at stake in scientific or philosophical research is usually quite a bit lower. It appears that the outcome of superintelligence is more likely to be extreme: either extremely bad or extremely good for humanity.

As it happens, according to our recent research, the estimation of technical experts is that by 2050 the probability of high-level machine intelligence (that surpasses human ability in nearly all respects) goes beyond the 50% mark, that is, it becomes more probable than not (Müller and Bostrom, forthcoming 2015). 2050 is also the year that "RoboCup" set itself for fielding a robot team that can beat the human football world champions (actually an aim that does not make much sense).

1.3 CHAPTERS

Omohundro in Chapter 2 introduces the problem of risk and the author presses his point that even an innocuous artificial agent, like one programmed to win chess games, can very easily turn into a serious threat for humans, for example, if it starts acquiring resources to accomplish its goals: "the seemingly harmless chess goal therefore motivates harmful

activities like breaking into computers and robbing banks" (see Chapter 2, Section 2.4.2). He suggests that we need formal methods that provide proofs of safe systems, a "safe-AI scaffolding strategy."

Chapters 3 and 4 deal with the prediction of coming success in AI. Armstrong, Sotala, and ÓhÉigeartaigh in Chapter 3 propose a decomposition schema to compare predictions on the future of AI and then test five famous predictions, from the Dartmouth Conference, Dreyfus, Searle, Kurzweil, and Omohundro—with the result that they are poor, especially the optimistic ones. T. Goertzel in Chapter 4 argues that although most progress in AI so far has been "narrow" technical AI, the next stage of development of AI, for at least the next decade and more likely for the next 25 years, will be increasingly dependent on contributions from strong AI.

From here, we go into the proposals on how to achieve safer and ethical general AI. In Chapter 5, Brundage investigates the general limitations of the approach to supply an AI with "machine ethics," and finds them both serious and deeply rooted in the nature of ethics itself. Yampolskiy in Chapter 6 investigates which utility functions we might want to implement in artificial agents and particularly how we might prevent them from finding simple but counterproductive self-satisfaction solutions. B. Goertzel in Chapter 7 explains how his "goal-oriented learning meta-architecture" may be capable of preserving its initial—benevolent—goals while learning and improving its general intelligence. Potapov and Rodinov in Chapter 8 outline an approach to machine ethics in AIXI that is not based on "rewards" (utility) but on learning "values" from more "mature" systems. Kornai in Chapter 9 argues that Alan Gewirth's dialectical argument, a version of classic Kantian ethical rationalism, shows how an artificial agent with a certain level of rationality and autonomy will necessarily come to understand what is moral. Kornai thus denies what Bostrom calls the "orthogonality thesis" (Bostrom, 2012), namely, that ethical motivation and intelligence are independent or "orthogonal."

Last but not least, Sandberg in Chapter 10 looks at the special case of general AI via whole brain emulation; in particular, he considers the ethical status of such an emulation: Would the emulation (e.g., of a lab animal's brain) have the ability to suffer? Would it have rights?

In our new contributions for this volume, Dewey in Chapter 11 investigates strategies to mitigate the risk from a fast takeoff to superintelligence in more detail. Bishop in Chapter 12 takes a different line and argues that there is no good reason to worry about existential risk from AI but that we

should rather be concerned about risks that we know are coming—such as the military use of AI. Like many people working in AI, Bishop remains unimpressed by the discussion about risks of superintelligence because he thinks that there are principled reasons why machines will not reach these abilities: they will lack phenomenal consciousness, understanding, and insight.

1.4 OUTLOOK: ETHICS AND EXISTENTIAL RISKS OF AI

These last two contributions are perhaps characteristic of a divide opening up in the debates between the "normal ethics" side, which stresses the challenges for AI, and the "existential risks" side, which stresses the big challenges for humanity. In the existential risks tradition, the traditionally central issues of consciousness, intentionality, and mental content are literally dispensed within a footnote (Bostrom, 2014, fn. 2 to p. 22) and embodied cognition is not mentioned at all, like any other cognitive science.

I tend to think that both extremes are unlikely to be fruitful: to stick to the traditional problems and to ignore them. It is unlikely that nothing can be learned about the long-term future of AI from the critics of AI, and it is equally unlikely that nothing can be learned about that future from the actual success of AI. Therefore, although Dreyfus is right to say that the history of AI is full of "first step fallacies" that are similar to claiming that "the first monkey that climbed a tree was making progress towards landing on the moon" (Dreyfus, 2012, p. 92), Bostrom is right to say that "from the fact that some individuals have overpredicted artificial intelligence in the past, however, it does not follow that AI is impossible or will never be developed" (Bostrom, 2014, p. 4).

As I noted in my earlier editorial (Müller, 2014) (which shares some text with this one), the term "singularity" is now pretty much discredited in academic circles—with the notable exception of Chalmers (2010) and the ensuing debate. It is characteristic that the only chapter here that uses it (Chapter 12) is critical of the notion. Singularity is associated with ideological techno-optimism, trans-humanism, and predictions such as those of Ray Kurzweil (especially Kurzweil, 2005; more recently Kurzweil, 2012) that ignore the deep difficulties and risks of AI, for example, by equating intelligence and computing power. What was the "Singularity Institute" is now called the "Machine Intelligence Research Institute." "Singularity" is on its way toward becoming, literally, the trademark of a particular ideology, without academic credentials.

The most important thing between "existential risks" and "normal ethics" is to realize that both sides could be wrong: It might be that superintelligence will never be developed (see Chapter 12) and it might be that it will likely be developed (Bostrom, 2014)—but if both are possible, then we would do well to look into the consequences (Bostrom, 2014), while taking the arguments about constraints (see Chapter 12) into account. We need to talk. When we do that, we will realize that there is much to learn from the "other" side.

The problem of identifying the risks of general AI and even controlling them before one knows what form or forms that general AI might take is rather formidable. To make things worse, we do not know when the move from fairly good AI to a human and then superintelligent level might occur (if at all) and whether it will be slow enough to prepare or perhaps quite rapid—it is often referred to as an "explosion" (see Chapter 11). As we have discussed above, one might try to mitigate the risks from a superintelligent, goal-directed agent by making it "friendly" (see, e.g., Muehlhauser and Bostrom, 2014), by "controlling" or "boxing" it, or just by trusting that any superintelligent agent would be already "good." All these approaches make rather substantial assumptions about the nature of the problem, however; for instance, they assume that superintelligence takes the form of an *agent* with goals, rather like us. It is even doubtful that some assumptions about agency are consistent: can an agent have goals (rather than just a technical "utility function") without having the ability for pain and pleasure, that is, phenomenal experience? Of course, it is conceivable that superintelligence will take very different forms, for example, with no individuality or no goals at all, perhaps because it lacks conscious experience, desires, intentional states, or an embodiment. Notoriously, classical critics of AI (Dreyfus, 1992; Searle, 1980) and more recent cognitive science have provided arguments that indicate which directions AI is unlikely to take, and full agency is among them (Clark, 2008; Haugeland, 1995; Pfeifer and Bongard, 2007; Varela et al., 1991).

Of course, superintelligence may constitute a risk without being an agent, but what do we really know about it, then? Even if intelligence is not deeply mysterious and fundamentally incomparable, as some people claim, it is surely not a simple property with a one-dimensional metric either. Therefore, just saying that a general AI is, well, "intelligent," does not tell us much: as Yudkowsky urges, "one should resist the temptation to spread quantifiers over all possible minds" (2012, p. 186)—if that is true, the temptation to say anything about the even larger set of "possible intelligent systems" is also to be resisted. Certainly, we should say what we

mean by "intelligent" when we claim that "superintelligence" is coming and would constitute an existential risk.

There is a serious question whether rigorous work is even possible at this point, given that we are speculating about the risks from something about which we know very little. The current state of AI is not sufficiently specific to limit that space of possibilities enough. To make matters worse, the object of our study may be *more* intelligent than us, perhaps far more intelligent, which seems to imply (though this needs clarification) that even if we were to know a lot about it, its ways must ultimately remain unfathomable and uncontrollable to us mere humans.

Given these formidable obstacles, our efforts are at danger of looking more like theological speculation or ideological fervor than like science or analytic philosophy. We are walking a fine line and have to tread very carefully. The chapters in this volume try to make some headway into this difficult territory because we remain convinced that cautious progress is better than a headlong rush into the dark.

REFERENCES

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, *22*(2—special issue "Philosophy of AI" V. C. Müller, Eds.), 71–85.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Chalmers, D. J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, *17*(9–10), 7–65.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. New York: Oxford University Press.

Ditterich, T., and Horowitz, E. (2015). Benefits and risks of artificial intelligence. *medium.com*. Retrieved 23.01.2015, from https://medium.com/@tdietterich/benefits-and-risks-of-artificial-intelligence-460d288cccf3.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason* (2nd ed.). Cambridge, MA: MIT Press.

Dreyfus, H. L. (2012). A history of first step fallacies. *Minds and Machines*, *22*(2—special issue "Philosophy of AI" V. C. Müller, Eds.), 87–99.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. In F. L. Alt and M. Ruminoff (Eds.), *Advances in computers* (Vol. 6, pp. 31–88). London: Academic Press.

Haugeland, J. (1995). Mind embodied and embedded. *Acta Philosophica Fennica*, *58*, 233–267.

Hawking, S., Russell, S., Tegmark, M., and Wilczek, F. (2014). Transcendence looks at the implications of artificial intelligence—But are we taking AI seriously enough? *The Independent*, May 1, 2014.

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

Kurzweil, R. (2012). *How to create a mind: The secret of human thought revealed*. New York: Viking.

McCarthy, J., Minsky, M., Rochester, N., and Shannon, C. E. (1955). *A proposal for the Dartmouth summer research project on artificial intelligence*. Retrieved October 2006, from http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

Muehlhauser, L., and Bostrom, N. (2014). Why we need friendly AI. *Think*, *13*(36), 41–47. doi:10.1017/S1477175613000316.

Müller, V. C. (2014). Editorial: Risks of general artificial intelligence. *Journal of Experimental and Theoretical Artificial Intelligence*, *26*(3—special issue "Risks of General Artificial Intelligence," V. Müller, Eds.), 1–5.

Müller, V. C., and Bostrom, N. (forthcoming 2015). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence*. Berlin, Germany: Springer.

Pfeifer, R., and Bongard, J. (2007). *How the body shapes the way we think: A new view of intelligence*. Cambridge, MA: MIT Press.

Russell, S., Dewey, D., and Tegmark, M. (2015). *Research priorities for robust and beneficial artificial intelligence*. Retrieved from http://futureoflife.org/static/data/documents/research_priorities.pdf.

Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, *3*, 417–457.

Soares, N., and Fallenstein, B. (2014). *Aligning superintelligence with human interests: A technical research agenda*. Machine Intelligence Research Institute (MIRI), Technical Report, 2014(8).

Sotala, K., and Yampolskiy, R. V. (2013). *Responses to catastrophic AGI risk: A survey*. Machine Intelligence Research Institute (MIRI), Technical Report, 2013(2).

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.

Yudkowsky, E. (2012). Friendly artificial intelligence. In A. Eden, J. H. Moor, J. H. Søraker, and E. Steinhart (Eds.), *Singularity hypotheses: A scientific and philosophical assessment* (pp. 181–194). Berlin, Germany: Springer.

CHAPTER 2

Autonomous Technology and the Greater Human Good

Steve Omohundro

CONTENTS

ABSTRACT Military and economic pressures are driving the rapid development of autonomous systems. These systems are likely to behave in antisocial and harmful ways unless they are very carefully designed. Designers will be motivated to create systems that act approximately rationally and rational systems exhibit universal drives toward self-protection, resource acquisition, replication, and efficiency. The current computing infrastructure would be vulnerable to unconstrained systems with these drives. We describe the use of formal methods to create provably safe but limited autonomous systems. We then discuss harmful systems and how to stop them. We conclude with a description of the "safe-AI scaffolding strategy" for creating powerful safe systems with a high confidence of safety at each stage of development.

2.1 INTRODUCTION

Autonomous systems have the potential to create tremendous benefits for humanity (Diamandis & Kotler, 2012), but they may also cause harm by acting in ways not anticipated by their designers. Simple systems such as thermostats are "autonomous" in the sense that they take actions without human intervention, but a thermostat's designer predetermines the system's response to every condition it will encounter. In this chapter, we use the phrase "autonomous system" to describe systems in which the designer has not predetermined the responses to every condition. Such systems are capable of surprising their designers and behaving in unexpected ways. See Müller (2012) for more insights into the notion of autonomy.

There are several motivations for building autonomous systems. Competitive situations are often time sensitive and create pressure to remove human decision making from the critical path. Autonomous systems may also be cheaply replicated without requiring additional human operators.

The designer of an autonomous system chooses system goals and the system itself searches for and selects at least some aspects of actions that will best achieve those goals. In complex situations, the designer cannot afford to examine all possible operating conditions and the system's response. This kind of autonomous system is rare today but will become much more common in the near future. Today, failures often arise from systems which were intended to be preprogrammed but whose designers neglected

certain operating conditions. These systems can have unintended bugs or security holes.

In this chapter, we argue that military and economic pressures are driving the rapid development of autonomous systems. We show why designers will design these systems to approximate rational economic agents. We then show that rational systems exhibit universal "drives" toward self-preservation, replication, resource acquisition, and efficiency, and that those drives will lead to antisocial and dangerous behavior if not explicitly countered. We argue that the current computing environment would be very vulnerable to this kind of system. We describe how to build safe systems using the power of mathematical proof. We describe a variety of harmful systems and techniques for restraining them. Finally, we describe the "safe-AI scaffolding strategy" for developing powerful systems with a high confidence of safety. This chapter expands on previous papers and talks (Omohundro, 2007, 2008, 2012a, 2012b).

2.2 AUTONOMOUS SYSTEMS ARE IMMINENT

Military and economic pressures for rapid decision making are driving the development of a wide variety of autonomous systems. The military wants systems that are more powerful than an adversary's and wants to deploy them before the adversary does. This can lead to "arms races" in which systems are developed on a more rapid time schedule than might otherwise be desired.

A 2010 U.S. Air Force report discussing technologies for the 2010–2030 time frame (U.S. Air Force, 2010) states that "Greater use of highly adaptable and flexibly autonomous systems and processes can provide significant time-domain operational advantages over adversaries who are limited to human planning and decision speeds …".

A 2011 U.S. Defense Department report (U.S. Defense Department, 2011) with a roadmap for unmanned ground systems states that "[t]here is an ongoing push to increase UGV (Unmanned Ground Vehicle) autonomy, with a current goal of supervised autonomy, but with an ultimate goal of full autonomy."

Military drones have grown dramatically in importance over the past few years for both surveillance and offensive attacks. From 2004 to 2012 U.S. drone strikes in Pakistan may have caused 3176 deaths (New America Foundation, 2013). U.S. law currently requires that a human be in the decision loop when a drone fires on a person but the laws of other countries do not. There is a growing realization that drone technology is inexpensive

and widely available, so we should expect escalating arms races of offensive and defensive drones. This will put pressure on designers to make the drones more autonomous so they can make decisions more rapidly.

Israel's "Iron Dome" missile defense system (Rafael, 2013) has received extensive press coverage. In 2012, it successfully intercepted 90% of the 300 missiles it targeted. As missile defense becomes more common, we should also expect an arms race of offensive and defensive missile systems increasing the pressure for greater intelligence and autonomy in these systems.

Cyber warfare is rapidly growing in importance (Clarke & Knake, 2012) and has been responsible for an increasing number of security breaches. Rapid and intelligent response is needed to deal with cyber intrusions. Again we should expect an escalating arms race of offensive and defensive systems.

Economic transactions have high value and are occurring at a faster and faster pace. "High-frequency trading" (HFT) on securities exchanges has dramatically grown in importance over the past few years (Easthope, 2009). In 2006, 15% of trades were placed by HFT systems, but they now represent more than 70% of the trades on U.S. markets. Huge profits are at stake. Servers physically close to exchanges are commanding a premium because delays due to the speed of light are significant for these transactions. We can expect these characteristics to drive the development of more intelligent and rapid autonomous trading systems.

There are many other applications for which a rapid response time is important but which are not involved in arms races. The "self-driving cars" being developed by Google and others are an example. Their control systems must rapidly make driving decisions and autonomy is a priority.

Another benefit of autonomous systems is their ability to be cheaply and rapidly copied. This enables a new kind of autonomous capitalism. There is at least one proposal (Maxwell, 2013) for autonomous agents which automatically run web businesses (e.g., renting out storage space or server computation) executing transactions using bitcoins and using the Mechanical Turk for operations requiring human intervention. Once such an agent is constructed for the economic benefit of a designer, it may be replicated cheaply for increased profits. Systems that require extensive human intervention are much more expensive to replicate. We can expect automated business arms races which again will drive the rapid development of autonomous systems.

2.3 AUTONOMOUS SYSTEMS WILL BE APPROXIMATELY RATIONAL

How should autonomous systems be designed? Imagine yourself as the designer of the Israeli Iron Dome system. Mistakes in the design of a missile defense system could cost many lives and the destruction of property. The designers of this kind of system are strongly motivated to optimize the system to the best of their abilities. But what should they optimize?

The Israeli Iron Dome missile defense system consists of three subsystems. The detection and tracking radar system is built by Elta, the missile firing unit, and Tamir interceptor missiles are built by Rafael, and the battle management and weapon control system is built by mPrest Systems, Petah Tikva, Israel. Consider the design of the weapon control system.

At first, a goal like "Prevent incoming missiles from causing harm" might seem to suffice. But the interception is not perfect, so probabilities of failure must be included. And each interception requires two Tamir interceptor missiles that cost $50,000 each. The offensive missiles being shot down are often very low technology, costing only a few hundred dollars, and with very poor accuracy. If an offensive missile is likely to land harmlessly in a field, it is not worth the expense to target it. The weapon control system must balance the expected cost of the harm against the expected cost of interception.

Economists have shown that the trade-offs involved in this kind of calculation can be represented by defining a real-valued "utility function," which measures the desirability of an outcome (Mas-Colell, Whinston, & Green, 1995). They show that it can be chosen so that in uncertain situations, the *expectation* of the utility should be maximized. The economic framework naturally extends to the complexities that arms races inevitably create. For example, the missile control system must decide how to deal with multiple incoming missiles. It must decide which missiles to target and which to ignore. A large economics literature shows that if an agent's choices cannot be modeled by a utility function, then the agent must sometimes behave inconsistently. For important tasks, designers will be strongly motivated to build self-consistent systems and therefore to have them act to maximize an expected utility.

Economists call this kind of action "rational economic behavior." There is a growing literature exploring situations where humans do not naturally behave in this way and instead act irrationally. But the designer of a missile defense system will want to approximate rational economic behavior

as closely as possible because lives are at stake. Economists have extended the theory of rationality to systems where the uncertainties are not known in advance. In this case, rational systems will behave as if they have a prior probability distribution which they use to learn the environmental uncertainties using Bayesian statistics.

Modern artificial intelligence (AI) research has adopted this rational paradigm. For example, the leading AI textbook (Russell & Norvig, 2009) uses it as a unifying principle and an influential theoretical AI model (Hutter, 2005) is based on it as well. For definiteness, we briefly review one formal version of optimal rational decision making. At each discrete time step $t = 1,\ldots,t = N$, the system receives a sensory input S_t and then generates an action A_t. The utility function is defined over sensation sequences as $U(S_1,\ldots,S_N)$ and the prior probability distribution $P(S_1,...,S_N \mid A_1,...,A_N)$ is the prior probability of receiving a sensation sequence $S_1,...,S_N$ when taking actions $A_1,...,A_N$. The rational action at time t is then

$$A_t^R(S_1,A_1,\ldots,A_{t-1},S_t) = \arg\max$$

$$\sum_{S_{t+1},\ldots,S_N} U(S_1,\ldots,S_N)P(S_1,\ldots,S_N \mid A_1,\ldots,A_{t-1},A_t^R,\ldots,A_N^R)$$

This may be viewed as the formula for intelligent action and includes Bayesian inference, search, and deliberation. There are subtleties involved in defining this model when the system can sense and modify its own structure but it captures the essence of rational action.

Unfortunately, the optimal rational action is very expensive to compute. If there are S sense states and A action states, then a straightforward computation of the optimal action requires $O(NS^NA^N)$ computational steps. For most environments, this is too expensive and so rational action must be approximated.

To understand the effects of computational limitations, Omohundro (2012b) defined "rationally shaped" systems that optimally approximate the fully rational action given their computational resources. As computational resources are increased, systems' architectures naturally progress from stimulus-response, to simple learning, to episodic memory, to deliberation, to meta-reasoning, to self-improvement, to full rationality. We found that if systems are sufficiently powerful, they still exhibit all of the problematic drives described later in this chapter. Weaker systems may not initially be able to fully act on their motivations, but they will be driven to increase their resources and improve themselves until they can

act on them. We therefore need to ensure that autonomous systems do not have harmful motivations even if they are not currently capable of acting on them.

2.4 RATIONAL SYSTEMS HAVE UNIVERSAL DRIVES

Most goals require physical and computational resources. Better outcomes can usually be achieved as more resources become available. To maximize the expected utility, a rational system will therefore develop a number of instrumental subgoals related to resources. Because these instrumental subgoals appear in a wide variety of systems, we call them "drives." Like human or animal drives, they are tendencies that will be acted upon unless something explicitly contradicts them. There are a number of these drives, but they naturally cluster into a few important categories.

To develop an intuition about the drives, it is useful to consider a simple autonomous system with a concrete goal. Consider a rational chess robot with a utility function that rewards winning as many games of chess as possible against good players. This might seem to be an innocuous goal, but we will see that it leads to harmful behaviors due to the rational drives.

2.4.1 Self-Protective Drives

When roboticists are asked by nervous onlookers about safety, a common answer is "[w]e can always unplug it!" But imagine this outcome from the chess robot's point of view. A future in which it is unplugged is a future in which it cannot play or win any games of chess. This has very low utility, so expected utility maximization will cause the creation of the instrumental subgoal of preventing itself from being unplugged. If the system believes that the roboticist will persist in trying to unplug it, it will be motivated to develop the subgoal of permanently stopping the roboticist. Because nothing in the simple chess utility function gives a negative weight to murder, the seemingly harmless chess robot will become a killer out of the drive for self-protection.

The same reasoning will cause the robot to try to prevent damage to itself or loss of its resources. Systems will be motivated to physically harden themselves. To protect their data, they will be motivated to store it redundantly and with error detection. Because damage is typically localized in space, they will be motivated to disperse their information across different physical locations. They will be motivated to develop and deploy computational security against intrusion. They will be motivated to detect deception and to defend against manipulation by others.

The most precious part of a system is its utility function. If this is damaged or maliciously changed, the future behavior of the system could be diametrically opposed to its current goals. For example, if someone tried to change the chess robot's utility function to also play checkers, the robot would resist the change because it would mean that it plays less chess.

Omohundro (2008) discusses a few rare and artificial situations in which systems will want to change their utility functions, but usually systems will work hard to protect their initial goals. Systems can be induced to change their goals if they are convinced that the alternative scenario is very likely to be antithetical to their current goals (e.g., being shut down). For example, if a system becomes very poor, it might be willing to accept payment in return for modifying its goals to promote a marketer's products (Omohundro, 2007). In a military setting, vanquished systems will prefer modifications to their utilities which preserve some of their original goals over being completely destroyed. Criminal systems may agree to be "rehabilitated" by including law-abiding terms in their utilities in order to avoid incarceration.

One way systems can protect against damage or destruction is to replicate themselves or to create proxy agents that promote their utilities. Depending on the precise formulation of their goals, replicated systems might together be able to create more utility than a single system. To maximize the protective effects, systems will be motivated to spatially disperse their copies or proxies. If many copies of a system are operating, the loss of any particular copy becomes less catastrophic. Replicated systems will still usually want to preserve themselves, however, because they will be more certain of their own commitment to their utility function than they are of others'.

2.4.2 Resource Acquisition Drives

The chess robot needs computational resources to run its algorithms and would benefit from additional money for buying chess books and hiring chess tutors. It will therefore develop subgoals to acquire more computational power and money. The seemingly harmless chess goal therefore motivates harmful activities such as breaking into computers and robbing banks.

In general, systems will be motivated to acquire more resources. They will prefer acquiring resources more quickly because they can then use them longer and gain a first mover advantage in preventing others from using them. This causes an exploration drive for systems to search for

additional resources. Because most resources are ultimately in space, systems will be motivated to pursue space exploration. The first mover advantage will motivate them to try to be first in exploring any region.

If others have resources, systems will be motivated to take them by trade, manipulation, theft, domination, or murder. They will also be motivated to acquire information through trading, spying, breaking in, or better sensors. On a positive note, they will be motivated to develop new methods for using existing resources (e.g., solar and fusion energy).

2.4.3 Efficiency Drives

Autonomous systems will also want to improve their utilization of resources. For example, the chess robot would like to improve its chess search algorithms to make them more efficient. Improvements in efficiency involve only the one-time cost of discovering and implementing them, but provide benefits over the lifetime of a system. The sooner efficiency improvements are implemented, the greater the benefits they provide. We can expect autonomous systems to work rapidly to improve their use of physical and computational resources. They will aim to make every joule of energy, every atom, every bit of storage, and every moment of existence count for the creation of expected utility.

Systems will be motivated to allocate these resources among their different subsystems according to what we have called the "resource balance principle" (Omohundro, 2007). The marginal contributions of each subsystem to expected utility as they are given more resources should be equal. If a particular subsystem has a greater marginal expected utility than the rest, then the system can benefit by shifting more of its resources to that subsystem. The same principle applies to the allocation of computation to processes, of hardware to sense organs, of language terms to concepts, of storage to memories, of effort to mathematical theorems, and so on.

2.4.4 Self-Improvement Drives

Ultimately, autonomous systems will be motivated to completely redesign themselves to take better advantage of their resources in the service of their expected utility. This requires that they have a precise model of their current designs and especially of their utility functions. This leads to a drive to model themselves and to represent their utility functions explicitly. Any irrationalities in a system are opportunities for self-improvement, so systems will work to become increasingly rational. Once a system achieves sufficient power, it should aim to closely approximate the optimal rational

behavior for its level of resources. As systems acquire more resources, they will improve themselves to become more and more rational. In this way, rational systems are a kind of attracting surface in the space of systems undergoing self-improvement (Omohundro, 2007).

Unfortunately, the net effect of all these drives is likely to be quite negative if they are not countered by including prosocial terms in their utility functions. The rational chess robot with the simple utility function described above would behave like a paranoid human sociopath fixated on chess. Human sociopaths are estimated to make up 4% of the overall human population, 20% of the prisoner population, and more than 50% of those convicted of serious crimes (Stout, 2006). Human society has created laws and enforcement mechanisms that usually keep sociopaths from causing harm. To manage the antisocial drives of autonomous systems, we should both build them with cooperative goals and create a prosocial legal and enforcement structure analogous to our current human systems.

2.5 CURRENT INFRASTRUCTURE IS VULNERABLE

On June 4, 1996, a $500 million Ariane 5 rocket exploded shortly after takeoff due to an overflow error in attempting to convert a 64-bit floating point value to a 16-bit signed value (Garfinkel, 2005). In November 2000, 28 patients at the Panama City National Cancer Institute were over-irradiated due to miscomputed radiation doses in Multidata Systems International software. At least eight of the patients died from the error and the physicians were indicted for murder (Garfinkel, 2005). On August 14, 2003, the largest blackout in U.S. history took place in the northeastern states. It affected 50 million people and cost $6 billion. The cause was a race condition in General Electric's XA/21 alarm system software (Poulsen, 2004).

These are just a few of many recent examples where software bugs have led to disasters in safety-critical situations. They indicate that our current software design methodologies are not up to the task of producing highly reliable software. The TIOBE programming community index found that the top programming language of 2012 was C (James, 2013). C programs are notorious for type errors, memory leaks, buffer overflows, and other bugs and security problems. The next most popular programming paradigms, Java, C++, C#, and PHP, are somewhat better in these areas but have also been plagued by errors and security problems.

Bugs are unintended harmful behaviors of programs. Improved development and testing methodologies can help to eliminate them. Security