MACHINE ETHICS AND ROBOT ETHICS

THE LIBRARY OF ESSAYS ON THE ETHICS OF EMERGING TECHNOLOGIES

Series editor: Wendell Wallach

Titles in the series

The Ethical Challenges of Emerging Medical Technologies
Edited by Arthur L. Caplan and Brendan Parent

The Ethics of Biotechnology
Edited by Gaymon Bennett

The Ethics of Nanotechnology, Geoengineering, and Clean Energy
Edited by Andrew Maynard and Jack Stilgoe

The Ethics of Sports Technologies and Human Enhancement
Edited by Thomas H. Murray and Voo Teck Chuan

Emerging Technologies
Edited by Gary E. Marchant and Wendell Wallach

The Ethics of Information Technologies
Edited by Keith W. Miller and Mariarosaria Taddeo

Machine Ethics and Robot Ethics
Edited by Wendell Wallach and Peter Asaro

The Ethics of Military Technology
Edited by Braden Allenby

Machine Ethics and Robot Ethics

WENDELL WALLACH,
Yale University, USA,
AND PETER ASARO,
New School for Public Engagement, USA

THE LIBRAY OF ESSAYS ON THE ETHICS OF EMERGING TECHNOLOGIES

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2017
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
A catalog record for this book has been requested

ISBN: 978-1-4724-3039-7 (hbk)

Typeset in Times New Roman MT
by Servis Filmsetting Ltd, Stockport, Cheshire

Contents

Part I: Laying foundations

Part II: Robot ethics

Part III: Machine ethics

Part IV: Moral agents and agency

Part V: Law and policy

Acknowledgments

The chapters in this volume are taken from the sources listed below. The editor and publishers wish to thank the authors, original publishers, or other copyright holders for permission to use their material as follows.

Chapter 1: Clarke, R. (1993). "Asimov's Laws of Robotics: Implications for Information Technology (1)." *IEEE Computer*, 26(12), 53–61. Permission from IEEE Computer Society.

Chapter 2: Clarke, R. (1994). "Asimov's Laws of Robotics: Implications for Information Technology (2)." *IEEE Computer*, 227(1), 57–66. Permission from IEEE Computer Society.

Chapter 3: Allen, C., Varner, G., & Zinser, J. (2000). "Prolegomena to Any Future Artificial Moral Agent." *Journal of Experimental and Theoretical Artificial Intelligence*, 12, 251–261. Permission from Taylor and Francis.

Chapter 4: Nissenbaum, H. (2001). "How Computer Systems Embody Values." *Computer*, 34(3), 118–119. Permission from IEEE Computer Society.

Chapter 5: Bostrom, N. (2003). "The Ethical Issues of Advanced Artificial Intelligence." Paper presented at the IIAS 2003, Baden Baden, GE. In Smit, S., Wallach, W., and Lasker, L. (eds.) *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol 11, IIAS, pp. 12–17.

Chapter 6: Veruggio, G., & Operto, F. (2006). "Roboethics: a Bottom-up Interdisciplinary Discourse in the Field of Applied Ethics in Robotics." *International Review of Information Ethics*, 6, 2–8. Permission from the International Review of Information Ethics.

Chapter 7: Asaro, P. (2006). "What Should We Want from a Robot Ethic?" *International Review of Information Ethics*, 6, 10–16. Permission from the International Review of Information Ethics.

Chapter 8: Sparrow, Robert (2004). "The Turing Triage Test." *Ethics and Information Technology* 6.4, 203–213. Permission from Springer.

Chapter 9: Turkle, Sherry (2006). "A Nascent Robotics Culture: New Complicities for Companionship." *American Association for Artificial Intelligence AAAI.* Permission from the American Association for Artificial Intelligence.

Chapter 10: Coeckelbergh, Mark (2010). “Moral Appearances: Emotions, Robots, and Human Morality.” *Ethics and Information Technology*, 12.3, 235–241. Permission from Springer.

Chapter 11: Borenstein, Jason, and Yvette Pearson (2010). “Robot Caregivers: Harbingers of Expanded Freedom for All?” *Ethics and Information Technology*, 12.3, 277–288. Permission from Springer.

Chapter 12: Vallor, Shannon (2011). “Carebots and Caregivers: Sustaining the Ethical Ideal of Care in the Twenty-first Century.” *Philosophy & Technology*, 24.3, 251–268. Permission from Springer.

Chapter 13: Sharkey, Noel, & Amanda Sharkey (2010). “The Crying Shame of Robot Nannies: An Ethical Appraisal.” *Interaction Studies*, 11.2: 161–190. Permission from John Benjamins Publishing Company.

Chapter 14: van Wynsberghe, Aimee (2013). “Designing Robots for Care: Care Centered Value-sensitive Design.” *Science and Engineering Ethics*, 19.2, 407–433.

Chapter 15: Sullins, John P. (2012). “Robots, Love, and Sex: The Ethics of Building a Love Machine.” *Affective Computing, IEEE Transactions*, 3.4, 398–409. Permission from the Institute of Electrical and Electronics Engineers.

Chapter 16: Malle, B., & Matthias Scheutz (2014). “Moral Competence in Social Robots.” *IEEE International Symposium on Ethics in Engineering, Science, and Technology, Chicago*. Permission from the Institute of Electrical and Electronics Engineers.

Chapter 17: Moor, James H. (2006). “The Nature, Importance, and Difficulty of Machine Ethics.” *Intelligent Systems, IEEE*, 21.4, 18–21. Permission from the Institute of Electrical and Electronics Engineers.

Chapter 18: Anderson, Michael, & Susan Leigh Anderson (2007). “Machine Ethics: Creating an Ethical Intelligent Agent.” *AI Magazine*, 28.4, 15–26. Permission from the AAAI Press.

Chapter 19: Wallach, Wendell, Colin Allen, & Iva Smit (2008). “Machine Morality: Bottom-up and Top-down Approaches for Modelling Human Moral Faculties.” *Ai & Society*, 22.4, 565–582. Permission from Springer.

Chapter 20: McDermott, Drew (2000). “Why Ethics is a High Hurdle for AI.” *North American Conference on Computing and Philosophy*. Bloomington, Indiana. Permission from the author.

Chapter 21: Powers, Thomas M. (2006). “Prospects for a Kantian Machine.” *Intelligent Systems, IEEE*, 21.4, 46–51.

Chapter 22: Guarini, Marcello (2005). “Particularism and Generalism: How AI Can Help Us to Better Understand Moral Cognition.” *Machine Ethics: Papers from the 2005 AAAI Fall symposium*. Permission from the AAAI Press.

Chapter 23: Bringsjord, S., Arkoudas, K., & Bello, P. (2006). “Toward a General Logicist Methodology for Engineering Ethically Correct Robots.” *IEEE Intelligent Systems*, 21(4), 38–44. Permission from the Institute of Electrical and Electronics Engineers.

Chapter 24: Wallach, Wendell, Colin Allen, & Stan Franklin (2011). “Consciousness and Ethics: Artificially Conscious Moral Agents.” *International Journal of Machine Consciousness*, 3.01, 177–192. Permission from the World Scientific Publishing Company.

Chapter 25: Floridi, Luciano, & Jeff W. Sanders (2004). “On the Morality of Artificial Agents.” *Minds and Machines*, 14.3, 349–379. Permission from Springer.

Chapter 26: Johnson, Deborah G., & Keith W. Miller (2008). “Un-making Artificial Moral Agents.” *Ethics and Information Technology*, 10.2–3, 123–133. Permission from Springer.

Chapter 27: Suchman, Lucy. “Agencies in Technology Design: Feminist Reconfigurations.” Hackett, Edward J., Olga Amsterdamska, Michael E. Lynch, and Judy Wajcman (eds.) *The Handbook of Science and Technology Studies,* third edition, excerpt from pp. 139–163, copyright 2007 Massachusetts Institute of Technology, by permission of The MIT Press.

Chapter 28: Marino, Dante, & Guglielmo Tamburrini (2006). “Learning Robots and Human Responsibility.” *International Review of Information Ethics*, 6, 46–51. Permission from the International Review of Information Ethics.

Chapter 29: Torrance, Steve (2014). “Artificial Consciousness and Artificial Ethics: Between Realism and Social Relationism.” *Philosophy & Technology*, 27.1, 9–29. Permission from Springer.

Chapter 30: Murphy, Robin R., & David D. Woods (2009). “Beyond Asimov: The Three Laws of Responsible Robotics.” *Intelligent Systems, IEEE*, 24.4, 14–20. Permission from the Institute of Electrical and Electronics Engineers.

Chapter 31: Solum, Lawrence (1992). “Legal Personhood for Artificial Intelligences.” *North Carolina Law Review*, 70, 1231–1287. Permission from the North Carolina Law Review.

Chapter 32: Nagenborg, Michael, et al. (2008). “Ethical Regulations on Robotics in Europe.” *Ai & Society*, 22.3, 349–366. Permission from Springer.

Chapter 33: Calo, M. Ryan (2010). “Robots and Privacy.” *Robot Ethics: The Ethical and Social Implications of Robotics*, 187–204. Permission from the MIT Press.

Chapter 34: Lin, Patrick (2014). "The Robot Car of Tomorrow May Just Be Programmed to Hit You." *Wired Magazine*, May 6, 2014.

Chapter 35: Gunkel, David J. (2014) "A Vindication of the Rights of Machines." *Philosophy & Technology*, 27, 113–132. Permission from Springer.

Every effort has been to trace all the copyright holders, but if any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangement at the first opportunity.

Publisher's note

The material in this volume has been reproduced using the facsimile method. This means we can retain the original pagination to facilitate easy and correct citation of the original essays. It also explains the variety of typefaces, page layouts, and numbering.

Series Preface

Scientific discovery and technological innovation are producing, and will continue to generate, a truly broad array of tools and techniques, each of which offers benefits while posing societal and ethical challenges. These emerging technologies include (but are not limited to) information technology, genomics, biotechnology, synthetic biology, nanotechnology, personalized medicine, stem cell and regenerative medicine, neuroscience, robotics, and geoengineering. The societal and ethical issues, which arise within those fields, go beyond safety and traditional risks such as public health threats and environmental damage, to encompass privacy, fairness, security, and the acceptability of various forms of human enhancement. The "Library of essays on the ethics of emerging technologies" demonstrates the breadth of the challenges and the difficult tradeoffs entailed in reaping the benefits of technological innovation while minimizing possible harms.

Editors selected for each of the eight volumes are leaders within their respective fields. They were charged to provide a roadmap of core concerns with the help of an introductory essay and the careful selection of articles that have or will play an important role in ongoing debates. Many of these articles can be thought of as "golden oldies," important works of scholarship that are cited time and again. Other articles selected address cutting-edge issues posed by synthetic organisms, cognitive enhancements, robotic weaponry, and additional technologies under development.

In recent years, information technologies have transformed society. In the coming decades, advances in genomics and nanotechnologies may have an even greater impact. The pathways for technological progress are uncertain as new discoveries and convergences between areas of research afford novel, and often unanticipated, opportunities. However, the determination of which technological possibilities are being realized or probable, and which are merely plausible or highly speculative, functions as a central question that cuts across many fields. This in turns informs which ethical issues being raised warrant immediate attention. Calls for precautionary measures to stave off harms from speculative possibilities can unnecessarily interfere with innovation. On the other hand, if futuristic challenges, such as smarter-than-human robotic are indeed possible, then it behooves us to invest now in means to ensure artificial intelligence will be provably beneficial, robust, safe, and controllable.

Most of the ethical concerns discussed in the volumes are less dramatic, but just as intriguing. What criteria must be met before newly created organisms can be released outside of a laboratory? Should fears about the possible toxicity of a few unidentified nanomaterials, among thousands, significantly slow the pace of development in a field that promises great rewards? Does medical research that mines large databases (big data), including the genomes of millions of people, have a downside? Are geoengineering technologies for managing climate change warranted, or more dangerous than the problem they purport to solve?

The ethical languages enlisted to evaluate an innovative technological go beyond the utilitarian analysis of costs and benefits. For example, the principles of biomedical ethics and the laws of armed conflict play a central role in judgments made about whether the healthcare industry or the military should adopt a proposed device or procedure. The differing ethical languages underscore different considerations, each of which will need to be factored into decisions regarding whether to embrace, regulate, or reject the new technology.

Scientific discovery and technological innovation proceed at an accelerating pace, while attention to the ethical, societal, and governance concerns they raise lags far behind. Scholars have been diligent in trying to bring those concerns to the fore. But, as the essays in these volumes will make clear, there is a great deal of work ahead if we, humanity as a whole, are to successfully navigate the promise and perils of emerging technologies.

Wendell Wallach

Introduction

The Emergence of Robot Ethics and Machine Ethics

Peter Asaro and Wendell Wallach

Once the stuff of science fiction, recent progress in artificial intelligence, robotics, and machine learning has raised public concern and academic interest in the safety and ethics of the new tools and techniques emerging from these fields of research. The technologies finally are coming into widespread use, and appear poised for rapid technological advancement. Some have even argued that the coming robotics revolution could rival the personal computer revolution of the 1970s and 1980s (Gates 2007), or the impact of the smartphone in the early twenty-first century.

As these technologies advance and find use in everyday life, they also raise a host of social, political, and legal issues. When artificial agents within computer systems and robots (hereafter collectively referred to as robots) start to act in the world—the physical world as well as in cyberspace—who is responsible for their actions? Should robots be used for the care of children or the homebound and the elderly? How can we ensure the safety of self-driving cars in a world of unpredictable pedestrians, playing children, dogs, bicyclists, and busy parking lots? How should we design robots to act, especially in cases with competing or conflicting goals? Can robots be designed so that they are sensitive to the values of those whom they interact with and that give form to the contexts in which they act? How do we design them to conform to our social and legal expectations? Can robots make moral decisions? Do we want them to make moral decisions? Then there are the more futuristic questions. Could computational systems be held accountable for actions they might initiate? How can we know if they have, or might eventually have, emotions or consciousness, or if they are moral agents deserving of rights and moral respect?

Speculation and science fiction have played an inordinately large, but not always welcomed, role in framing the ethical challenges posed by artificial intelligence and robotics. While some focus upon more speculative futures, as robots become reality and are introduced into the commerce of daily life, others are concerned with the more immediate practical challenges of ensuring their activity is safe and conforms to human laws and ethical considerations.

This volume is a collection of, and introduction to, scholarly work that focuses upon robots and ethics. The topic is divided into two broad fields of research. Robot ethics or roboethics explores how people should design, deploy, and treat robots (Veruggio & Operto 2006). It is particularly interested in how the introduction of robots will change human social interactions and what human social concerns tell us about how robots should be designed. Machine ethics or machine morality considers the prospects for creating computers and robots capable of making explicit moral decisions (Allen, Varner, & Zinser 2000; Moor 2006; Anderson & Anderson 2007). What capabilities will increasingly autonomous robots require: 1) to recognize when they are in ethically significant situations, and 2) to factor human ethical concerns into selecting safe, appropriate, and moral courses of action? There is no firm distinction between robot ethics and machine ethics, and some scholars treat machine ethics as a subset of robot ethics. Many of the scholars represented feel their work draws upon and contributes to both fields.

Machine ethics and robot ethics are both emergent intellectual disciplines—research fields which are only beginning to take form. The articles we have selected for reproduction are largely essays of historical significance that have become foundational for research in these two new fields of study. We sought to both convey the range of issues addressed and to include representative essays from leading figures in the development of robot ethics and machine ethics. The emphasis on a historical perspective is somewhat at the expense of more recent contributions. However, a few important recent articles are included. Indeed, a very recent shift in research on artificial intelligence and machine learning has directed much more attention by computer scientists to ethical concerns that had been largely of interest only to philosophers and social theorists.

This introduction opens with a discussion as to how robot ethics and machine ethics emerged out of earlier topics in science fiction, philosophy, cognitive science, and other fields of research. Then we turn to placing the articles included in the five sections of this book into context, before elucidating the shift in computer science that is directing much more attention to robot ethics and machine ethics. The sections are roughly, but not strictly, chronological in order to provide a sense of the development of various issues and positions. Ideas evolve over time through insights, intuitions, and scholarship put forward by countless individuals, only a few of whom are ever recognized.

The first section, "Laying foundations" attempts to capture a range of foundational work, which might be considered the "pre-history" of the fields. "Robot ethics" contains writings that established this area as a sub-discipline of technology ethics in its own right. Similarly, "Machine ethics" contains articles that established machine ethics as a distinct field. The last two sections focus on the two aspects of applied ethics—the internal perspective, the necessary conditions for considering a robot as an autonomous agent, and the external perspective of laws, regulations, and policies central for integrating robots into social life. In many cases, the issues, such as privacy, are relevant to both machine ethics and robot ethics, and are often relevant to larger questions in the ethics of emerging technologies more generally. The section on "Moral agents and agency" examines the ways in which robots and artificial software agents might be construed as, or designed to be, moral agents, and how this leads us to reconsider the nature of agency more generally. The section on "Law and policy" looks at the legal and policy-making issues we face in dealing with sophisticated robots. Advanced robotics will challenge many existing assumptions about what constitutes an agent and which tasks can be performed by technological systems.

The topics covered in this volume often overlap with those in companion volumes, particularly those in the volume on the *Ethics of Information Technologies*. Robots are essentially information systems that are mobile and can act in the world through remote control or autonomously. Furthermore, machine ethics encompasses ethical decision-making by not only robots, but also increasingly autonomous agents within computer networks. Broader issues in information and computer ethics, such as privacy, are largely covered in the volume on the *Ethics of Information Technologies*, while this volume goes into much greater depth discussing prospects for creating systems capable of making moral decisions. In addition, ethical issues related to the roboticization of warfare are covered in the volume on the *Ethics of Emerging Military Technologies*.

Finally, this introduction will outline recent advances in the development of artificial intelligence. Ensuring that intelligent systems will not only be safe, but also demonstrably beneficial and controllable is becoming an issue of much greater importance, and will continue to expand over the coming decade.

Science fiction and speculation

The speculative worlds of science fiction have given rise to imaginative technologies, fueled consumer desires, and guided the design of real products suited for everyday life. These worlds have also provided lessons in the ways that advanced systems might run amuck, and stoked fears over future technology.

While the public imagination is torn between the utopic and dystopic extremes of science fiction, engineers struggle to design limited purpose machines that will perform safely when introduced into the commerce of daily life. Some futurists proclaim the near-term advent of smarter-than-human robots that may or may not be friendly to our species (Kurzweil 2005). It is unfortunate, however, that this garners more media attention than real world problems such as the threat to privacy posed by the introduction of drones into civilian air space and the appropriateness of turning over aspects of elder care to robots.

Furthermore, speculation and science fiction feed misunderstandings regarding the capabilities of present day machines and those likely to be realized in the coming decade. In spite of the fact that IBM's Watson beat human champions on the TV show *Jeopardy*, robots lack higher-order cognitive capabilities, such as consciousness, semantic understanding, sophisticated learning abilities, discrimination, and empathy. Furthermore, the natural human tendency to anthropomorphize robots obfuscates recognition of their limitations. Today's robots are essentially machines. And yet they can be more skilled than their human counterparts at performing certain activities such as searching large databases, and revealing associations between disparate pieces of information.

Our challenge for this volume lies in grounding the discussion of robot ethics and machine ethics to focus on practical considerations that affect the way machines are designed and the tasks they might fulfill. Nevertheless, the three great themes of science fiction stories and films are helpful for framing this discussion.

1. Robots run amuck and cause harm—because they are dumb or rigid, because they were malevolently programmed, because they are alien or act in unanticipated ways, or because they wish to be free from servitude (*R.U.R.*, *Metropolis*)

2. Robots become social companions and even learn to apply ethical norms to determine how to act in new situations (Asimov's Laws, Cmd. Data in *Star Trek*)
3. Superintelligent forms of artificial intelligence that either seek to exterminate humans or become benevolent overlords (*2001: A Space Odyssey*, *The Terminator*, *The Matrix*, *Battlestar Galactica*)

Robot ethics is situated in a larger and older cultural history of robots in literature. From antiquity's Pygmalion to golems to Frankenstein to Rossum's Universal Robots, traditional narratives about humanoids are morality tales that emerge from clearly defined cultural fears. In nearly all the early science fiction, even initially benign robots became dangerous. They were usually emotionless, and often alien in origin, with little sympathy or concern for humans and the societal effects of their actions. Those robots that were not alien in origin were generally built by a scientific genius, whose robots eventually turn against them, if not all of humanity.

The implied moral of those stories is that scientific hubris leads to the building of monsters, and the eventual destruction of the scientist, monster, or both.

Isaac Asimov sought to break away from this traditional morality tale, and find one more compatible with his own brand of cautionary scientific optimism. He created the genre of the beneficent or good and helpful robot. In a series of short stories and novels, he proposed Three Laws that are meant to ensure robots act as helpful servants. The Three Laws (not harming humans, obeying humans, and self-preservation) are arranged hierarchically so that the first trumped the other two, and the second trumps the third. The Robot Laws were a precursor to machine ethics. But of course, what Asimov demonstrates in the course of exploring his Laws of Robotics is that they were flawed, inadequate, and undesirable in a surprising variety of circumstances. Later he added a fourth or Zeroth Law specifying that, "A robot may not harm humanity, or, by inaction, allow humanity to come to harm," but this only came into play in a few situations. The characters in his stories sometimes question and doubt the Laws' adequacy, and even when they do not, the readers do. In short, the Laws of Robotics are an excellent literary device, but wholly inadequate for the engineering and design of real robots and systems. Despite this, there are valuable lessons to be learned for machine and robot ethics by studying Asimov's stories and recognizing the nature of the failures to which they lead.

Integral to these literary lessons are questions of ethical value and morality. In arguing that we want artificial agents and robots to act in one way rather than another, we are expressing a desire, and with it a value. It is, in this regard, not unlike our expectations of humans, and our desire for them to act a certain way. We want people to be good, to act ethically, to reason morally, and to be virtuous. This is the stuff of moral theory, a Western philosophical tradition that was established by the ancient Greeks including Socrates, Plato, and Aristotle. That tradition has developed and elaborated various frameworks for describing and thinking about what constitutes the good life, moral actions, and the virtuous person.

This philosophical tradition has a long history of considering hypothetical and even magical situations (such as the Ring of Gyges in Plato's *Republic*, which bestows invisibility on its wearer). The nineteenth and twentieth centuries saw many of the magical powers of myth and literature brought into technological reality. Alongside the philosophical literature emerged literatures of horror, science fiction, and fantasy, which also explored many of the social, ethical, and philosophical questions raised by the new-found powers of technological capabilities. For most of the twentieth century, the examination of the ethical and moral implications of artificial intelligence and robotics was limited to the work of science fiction and cyberpunk writers, such as Isaac Asimov, Arthur C. Clarke, Bruce Sterling, William Gibson, and Philip K. Dick (to name only a few). It is only in the late twentieth century that we begin to see academic philosophers taking up these questions in a scholarly way.

Emergent disciplines

Like most new fields of knowledge, machine ethics and robot ethics emerged from work in other disciplines. Early interest came from obvious sources, such as engineering ethics and computer ethics, which were already attuned to the ethical and social value issues raised by new technologies. But contributions also came from other more developed subject areas in technology and ethics, such as bioethics and nanoethics. From within engineering came contributions from human–computer interaction and human–robot interaction. Taking a very practical approach to designing systems to interact with people, these fields became interested in the role of emotions in human social interactions, and how best to incorporate this knowledge into their designs. As practical fields, they already lay at an interdisciplinary crossroads of engineering, design, and social science.

Joining this mix were those philosophers and cognitive scientists interested in understanding the theoretical and empirical nature of human cognition, moral reasoning and moral intuitions, and decision-making. These researchers saw artificial agents and robots as experimental instruments or laboratories for developing simulations through which theories could be tested. With the help of simulations modeled within computational systems, they hoped to study and refine moral theories and better understand how humans reason when confronted with ethical dilemmas.

Some researchers came to the new fields out an interest in science fiction and cyberpunk literature, or a growing concern that science fiction technologies might become realities. Among these participants in the field, many are involved in the social movements concerned with human enhancements—known as transhumanism. They took an interest in robotic research and its potential for creating prosthetics that would enhance mind and body. Future robots might even serve as avatars into which one's mind could be uploaded and one's life extended indefinitely. While such concerns have not been central to either machine ethics or robot ethics, there are many points of contact and shared interests over the governance of emerging technologies and their capacity to transform how we think about what it means to be human.

These various disciplines and interests came together to form the highly interdisciplinary fields of machine ethics and robot ethics, and still continue to define the major lines of inquiry within the fields. The research has focused on four types of questions: the control and governance of computational systems; the exploration of ethical and moral theories using software and robots as laboratories or simulations; inquiry into the necessary requirements for moral agency and the basis and boundaries of rights; and questions of how best to design systems that are both useful and morally sound.

The legacy of bioethics and nanoethics can be seen most clearly in the research concerned primarily with the control and governance of artificial agents and robots as emergent technology. There are, of course, many emerging technologies in our contemporary world, each with their own ethical challenges, yet not all of these seem deserving of their own distinct academic fields of ethics. When a new technology is seen as potentially autonomous, or "other," that it might too easily get out of human control, then specialized governance mechanisms come into play. Similarly, when a new technology is so disruptive of existing technological regulations, then a new framework is needed. Such was the case with internet governance, or now with the introduction of unmanned aerial vehicles (drones) into civilian airspace, where whole new regimes of regulation must be imagined and implemented. Many of these are questions of social policy, but also questions of applied ethics.

Another powerful line of inquiry points to the meta-ethical questions. The potential for artificial moral agents to provide a laboratory of sorts for exploring the nature of moral theories computationally offers a meta-ethical research program. Which theories are computable or incomputable? What does it mean for practical determinations of right and wrong if ethics is not reducible to computable rules and prescriptions for behavior? What computations do humans actually use in selecting appropriate behavior in morally significant situations? How do we translate moral rules into computer code, and how do we interpret the results of following those moral rules? What does it mean to imitate, simulate, or replicate moral reasoning?

Related to these meta-ethical questions is a line of inquiry into the basis and boundaries of rights and the necessary requirements for moral agency. Can software or robotic agents become moral agents? What capabilities would they need for this? How would we know whether they have those capabilities? These in turn raise questions about how we recognize and accept other people as moral agents. Furthermore, artificial agents can take many forms that humans cannot, and thus can pose challenges to many straightforward analyses.

Finally, taking into account all of the above, how are we to implement practical machine ethics and robot ethics in everyday systems? As an engineer and designer, already engaged in developing a system that will interact with people and thus with the potential to help or do harm, how can one ensure a "good" design? From specific design principles, to design methodologies, to the application of domain knowledge, this is a rich and dynamic field. This work will hopefully grow in positive directions as the fields apply their research to the questions of how best to design systems that are both useful and morally sound.

Laying foundations

Questions that have become foundational for machine ethics and robot ethics began to intrigue scholars in the 1980s and 1990s. Sam Lehman-Wilzig (1981), in an article titled "Frankenstein Unbound: Towards a Legal Definition of Artificial Intelligence," questioned whether legal barriers exist that would preclude designating intelligent systems as accountable and liable for their actions. Roger Clarke (1993, 1994) was the first engineer to ask whether Asimov's Laws of Robotics could actually be implemented. He proposed that an engineer might well conclude that rules or laws are not an effective design strategy for building robots whose behavior must be appropriate, legal, and morally acceptable. Then in 1995, an edited volume entitled *Android Epistemology* contained two articles (not included in this volume), one by James Gips and another by Daniel Dennett, which raised critical issues. Gips considers whether ethical theories such as consequentialism and Kant's categorical imperative could be implemented within a computer system, and Dennett's chapter ponders the question in its title, "When Hal Kills, Who is to Blame?"

Debates around information and communication technologies (ICT) in the 1980s and 1990s were also seminal in the emergence of robot ethics and machine ethics. These debates emanated primarily from the experiences garnered from implementing large ICT systems within organizations. As such, there was a strong influence from the professional information technology system design field, as well as the more academic computer ethics field. Nissenbaum (2001) articulates the ways in which the design of computer systems involves innumerable judgments of value, and the implementation and use of any ICT system also enforces a set of values upon its users. The same can be said about any sophisticated machine, including robotic systems.

Allen, Varner, and Zinser (2000) go well beyond Gips' 1995 article in laying out the central challenges for creating artificial moral agents—disagreement among ethical theorists as to what the "right" ethical theory to implement is, and the computational challenges of implementing the various proposed theories. Much of the work in machine ethics follows their suggestion that practical issues of implementation—getting something that works, and that reduces the risks of a system causing harm through its actions—is more important than settling the philosophical disputes, or implementing a perfect morality. In addition, they consider whether a Moral Turing Test might be a means for establishing whether an artificial agent makes "good" decisions.

At the end of the section we return to the question which many people still feel belongs in the realm of science fiction, but which has also been raised more recently by an increasing number of leading scientists and technologists, including Stephen Hawking, Elon Musk, and Stuart Russell. Might computer scientists create an artificial intelligence that is not only smarter than us, but is capable of learning and growing at a much greater rate than humans, such that its working will be unknown to us and its capabilities unforeseeable and unimaginable? In an early paper, Bostrom (2003) outlines the kinds of ethical questions that arise in developing a superintelligence of this sort, and what it means for the human race to risk existential threats in order to derive what may be immeasurable benefits. The advent of superintelligence continues to be of ongoing interest to Bostrom, who organized reflections by himself and others on the subject into an influential 2014 book titled *Superintelligence: Paths, Dangers, Strategies.*

Robot ethics

Robot ethics emerged out of the broader field of engineering ethics and its subfield computer ethics. The initial impetus for robot ethics was concern over how to design robots that would not harm people, and thus shares its origins with both Asimov's Laws and the traditional engineering concerns directed at building tools that are safe. Safety alone was fine as an engineering approach for industrial robots, which operated in isolation from people, mostly in factories and often in safety cages. As robots became more capable, and it became apparent that robots would soon enter into the social world people inhabit, a whole range of possible harms and new design questions came to the fore.

"Roboethics," a term coined by Veruggio and Operto (2006), brings together the various concerns of applied engineering ethics in the context of robotics. They also explicitly include the range of ethical concerns raised in regard to ICT, and examine the development of codes of ethics in ICT that might be similarly deployed in robotics engineering. Asaro (2006) examines whether there might be a systematic and combined approach to robot ethics and machine ethics, which recognizes the range of robotic systems from their moral consequences to their moral agency.

A key aspect of robot ethics concerns the feelings and beliefs people hold towards robots. This includes a range of psychological and behavioral approaches that consider how much people identify with robots, or view them as having beliefs and feelings like people or perhaps like animals. Turkle (2006) examines various psychological relationships that people develop with their companion robots in clinical settings. Coeckelbergh (2010) considers the relationship between emotion and morality, and the implications of simulating emotional qualities in robots that are not capable of true emotions. Sparrow (2004) asks at what point might we decide to save a robot over a human life in a triage situation.

A central area of concern in robot ethics is their use in the domains of caregiving. While many human jobs do not apparently require essentially human traits or qualities to complete (assuming adequate machine intelligence and dexterity), caregiving appears to require uniquely human qualities. In particular, care that is provided to especially vulnerable populations, including children, the elderly, and the sick and injured, demands more humanity than other jobs. In caregiving there appears to be more to lose from automation and especially from poorly designed automation. Borenstein and Pearson (2010), Vallor (2011), Sharkey and Sharkey (2010) and Van Wynsburghe (2013) all address the question of robot caregivers from various perspectives and for various populations. Similar questions also arise in the development of robots designed for sexual love and companionship (Sullins 2012).

The section concludes with a paper that asks whether the moral competence of robots is sufficient for designing good robots (Malle and Schuetz 2014). That is, should robot engineers settle for robots that are competent in dealing appropriately with the moral issues they might encounter in their interactions with humans? Of course, being able to deal with moral issues means being able to interpret the moral relevance of situations and actions, making moral judgments, and communicating about morality. This article thus provides a nice segue to the field of machine ethics, and how we might actually design moral discernment and decision-making capabilities into robotic systems.

Machine ethics

The new field of research that is focused on the prospects for designing computer and robotic systems that demonstrate sensitivity to human values, and factor these into making decisions in morally significant situations, goes by various names. These include machine ethics, machine morality, moral machines, computational ethics, artificial morality, safe AI, and friendly AI. Each of these names gets used somewhat differently. Friendly AI, for example, refers explicitly to the challenge of ensuring smarter-than-human robots will be friendly to humans and human concerns, while safe AI encompasses many dimensions of robot safety, not merely whether robots can explicitly make moral decisions.

The essays in this section were selected primarily from early examples that broke new ground. Many of the very same scholars have discussed similar issues and additional experiments in fuller and even richer detail in their more recent papers. James Moor (2006) provided very useful distinction for helping to clarify when machines might actually be engaged in making explicitly moral decisions. Michael Anderson (a computer scientist) and Susan Leigh Anderson (a philosopher) performed some of the very first experiments in machine ethics, described in a 2007 article. Their computational system is confronted with a classic challenge in medical ethics over what to do when the human patient it delivers medicine to refuses to take the pills. Wallach, Allen, and Smit (2008) flesh out the distinction between bottom-up and top-down approaches to implementing moral decision-making faculties in robots and the need for capabilities beyond the ability to reason. A robot's choices and actions are generally hard wired, or what is sometimes referred to as *operational morality*. As systems become increasingly autonomous, the need to build in ethical routines for selecting appropriate behavior from among various courses of action (*functional morality)* will expand. But *functional morality* can still fall far short of full moral agency. These three articles introduce some of the most fundamental concepts and approaches that gave early form to machine ethics as a recognized field of study.

Robotics is a young field. To date, machine ethics has been dominated by philosophers and social theorists. Very few actual efforts to implement explicit moral decision-making capabilities in computer systems and physical robots have been made. Nevertheless, physical robots are entering the commerce of daily life. In 2008, Drew McDermott, a Professor of Computer Science at Yale University, weighed in as to not only why ethical decisions would be difficult for artificial intelligence systems to make, but also why he did not consider it a particularly important project for computer science at that time. Implicitly, McDermott's critique underscored why machine morality has been such a fascinating subject for philosophers, but of less interest to computer scientists.

A number of different scholars considered whether ethical theories such as utilitarianism or Kant's categorical imperative could actually be instantiated in a robot. The reproduced article by Tom Powers (2006) is but one representative of that sub-genre. The essay by Marcello Guarini (2005) describes the first of his many experiments to explore whether training neural networks with a degree of sensitivity to moral considerations might help us better understand human moral cognition.

There is a longstanding issue in moral philosophy as to whether ethics are, or should be, logically consistent. If ethics is logically consistent, then that might ease its implementation within a computational system. Bringsjord, Arkoudas, and Bello outline (2006) an approach for building robots that select “ethically correct” actions through deontic logic.

Consciousness is often discussed as an essential capacity for moral agency that might be difficult, if not impossible to implement in a robot. And yet, there is a field of research known as machine consciousness, which is attempting to design conscious robots. One of that field’s practitioners is Stan Franklin, a Professor at the University of Memphis, who is developing a conceptual and computational model for a robot (LIDA) with artificial general intelligence—the capability to solve a broad array of problems in many different domains. In early writings (Wallach & Allen 2008; Wallach, Franklin, & Allen 2010) the authors discuss how Franklin’s LIDA might be adapted to make moral decisions. Wallach, Allen, & Franklin (2011) go a step further in describing the functional role consciousness plays in ethical decision-making.

Collectively, the selected articles make it apparent why reflecting upon moral decision-making for robots has been such a valuable thought experiment for philosophers and scientists who wish to understand human cognition. Regardless of whether the field helps make future robots safer and more sensitive to ethical considerations, it has already injected vital and fruitful approaches for the better understanding of human ethical behavior.

Moral agents and agency

Even if we succeed in creating machines with moral sensitivity, there remains a separate question of whether their decisions and actions are those of moral agents, or mere outputs of a machine. Does the behavior of such machines actually constitute intentional agency? This section covers how work in the field has engaged the fundamental question of what constitutes moral agency as well as the more practical question of how to simulate or instantiate that capacity in an artificial agent. What constitutes agency can be construed in a variety of ways. Philosophers often approach agency as either a property requiring an essential quality or set of capabilities, or as something constituted through acts of self-determination. Legal scholars point to the conditions for the legal agency of people, corporations, and even animals. Engineers view agency as a matter of effective control, which can be passed between human operators and automatic systems. There is a narrow construal of agency that looks only at the properties of an individual agent, and broad perspectives that consider the social, political, and bureaucratic forces that usually shape the environment and decision contexts that individuals face. Some theorists take liberal views of agency and find it operating nearly everywhere, while others see it as a restricted class that should only be applied to appropriately sophisticated types of beings. An important distinction is made between moral patients and moral agents. Moral patients are those who are deserving of moral respect, for example, children, the mentally disabled, and some animals, despite not always being held fully responsible for their actions. Moral agents are those who are capable of moral action, moral responsibility, and showing moral respect to others.

A central tension in this subject area is the relationship between the agency of those who create artificial agents and the artificial agents themselves—and who is responsible, culpable, and accountable for the actions and consequences of using those systems. Floridi and Sanders (2004) propose a highly inclusive concept of moral agency, and articulate how they see artificial agents as being meaningfully considered as moral agents. Such agents need not have much intelligence at all, nor most of the qualities traditionally identified as essential to moral agency. Johnson and Miller (2008) directly challenge that approach to agency, and whether it can serve the role necessary for properly assigning responsibility, culpability, and accountability. They argue that such an approach can too easily mislead the identification of artificial agents as being responsible moral agents, when it is the human who designs and deploys them and is more appropriately seen as the responsible agent.

Suchman (2007) takes a broader and more critical analysis of the nature of agency in shaping the creation of artificial systems and the structuring of design processes themselves, challenging any easy solution to defining or replicating agency in artificial entities. This also calls into question applied frameworks for responsibility, culpability and accountability, as the application of these concepts is continually being politically contested and re-negotiated by participating parties, rather than being clear-cut concepts which agents readily conform to.

Marino and Tamburrini (2006) expand the challenges facing agency for artificial entities by considering the implications of machine learning and more open-ended learning systems. The more a learning machine adapts its programming based on data from its environment and its experiences, the less it is influenced by its initial programming. Does this imply that the human designers become less responsible over time, or that the environment and social forces, or the system itself, become more responsible? Torrance (2014) considers the nature of consciousness and its importance for determining artificial agency. What are the implications of treating consciousness, whatever we collectively determine it to be, as either an objectively real property of artificial agents, or as a product of social relativism?

In the final article for this section, Murphy and Woods (2009) take a position on agency which ducks the philosophical questions in favor of grounding the discussion in the nature of effective control, and the passing of such control among human operators and between human operators and automated systems. They do this through a critique and reformulation of Asimov's Three Laws of Robotics, viewed in terms of the responsiveness of robots to human operators, and the smooth transfer of control back and forth between the autonomous system and human operators. In focusing on control, they aim to make the shift in responsibility, and thus agency, clearer in dynamic and fluid situations.

Law and policy

It is the job of law and policy to articulate and enforce shared moral norms. In this section we examine how legal scholars have engaged the questions of the legal status of artificial agents, and their regulation. The question of legal agency or legal personhood for artificial agents is an extension of the agency debates covered in the previous section. However, it draws upon a large body of legal scholarship and real-world cases and laws on everything from the treatment of minor children, sports, and entertainment agents representing clients, intellectual property rights, privacy rights, damage resulting from the keeping of domesticated and wild animals, product liability, and the legal personhood of corporations. Many of these concepts can be applied to trying to predict how the courts will handle various near-term issues, such as those involving drones and self-driving cars. They also provide fertile analogies for thinking about more long-term issues such as when and whether machines and robots might have claims to legal and civil rights under the law.

Solum (1992) provides an in-depth analysis of what constitutes legal personhood, and how various aspects of the law might treat the actions of an artificial agent as being a legal person, potentially subject to prosecution or lawsuits. A more recent article by Asaro (2011) and a book by White & Chopra (2011) go into greater detail about the various forms of legal agency and how to interpret artificial software agents and robots, though we do not include articles by them in this section.

In terms of national policy for the regulation of robots, it is clear that ethics will play a serious role. It is less clear what form those regulations might take, or what governmental agencies might implement them. Nagenborg (2008) offers a perspective on how Europe might implement ethical regulations for robots.

One of the most significant ethical questions will be the privacy of information that robots are able to collect. A personal robot that has intimate knowledge of its owner might share that information with its manufacturer or with third parties, not unlike computer and smartphone technologies today, which already raise a host of privacy issues. It may also collect information about a variety of other individuals that it encounters or interacts with, quite unlike previous technologies. Calo (2010) considers these and other privacy issues raised by robots.

Self-driving cars present an example of a leading application of robotic technology that carries both serious risks and has a great potential to save lives. Lin (2014) considers the challenges to programming self-driving cars that may encounter situations where an accident is unavoidable. Who or what should the car effectively decide to hit in such situations?

Perhaps the most futuristic and contested question in robot ethics is whether robots may someday have the capacity or capability to be the bearer of rights, and in virtue of what criteria will they deserve those rights? Gunkel (2014) defends the view that there may be robots deserving of the legal and moral status of having rights, and challenges the current standards of evaluating which entities (corporations, animals, etc.) deserve moral standing. Such futuristic legal questions underscore the way in which artificial agents presently serve as a thought experiment for legal theorists and a vehicle for reflections on the nature of legal agency and rights.

Growing attention

As we completed work on this volume, there was tangible evidence of a sudden growth in the attention given to the ethical and societal challenges posed by artificial intelligence and robotics. Four different issues contribute to rising concerns as to whether the societal impact of robotics can be controlled in a manner that will minimize harms. First, the movement to ban Lethal Autonomous Weapons Systems (a.k.a. “Killer Robots,” see www.stopkillerrobots.org) has gained international attention and is now on the agenda of the Convention for Certain Conventional Weapons at the United Nations, to explore whether an arms control agreement might be forged. While both of us have been involved with that issue, it is not a subject discussed in this volume, but is covered in the companion volume on the *Ethics of Emerging Military Technologies*.

Second, the impact of robots on employment and wages is gaining attention with books and pronouncements by leading economists and social theorists. There is serious concern that the long-standing Luddite fear that technology might eliminate more jobs than it creates, may have finally come to pass. Robotics is not the only factor that contributes to what John Maynard Keynes (1930) named “technological unemployment.” But as artificial intelligence systems become increasingly intelligent and capable of performing more and more human jobs, their impact on the availability of jobs with satisfactory hourly wages will be immense.

Third, the proliferation of both large commercial and small hobby drones in civilian airspace and the advent of self-driving cars are garnering considerable attention as to the harms each might cause. Unusual ethical situations that self-driving cars might encounter have been discussed in articles by Gary Marcus (2012), Patrick Lin (2014, mentioned above and reproduced here), and others, and perhaps helps clarify why self-driving cars, heralded for years, are not yet being marketed. Furthermore, self-driving cars provide an apt metaphor for apprehensions regarding the trajectory of technological development. The incessant acceleration of innovation can appear to have put technology in the driver’s seat, both figuratively and literally, as the primary determinant of humanity’s destiny.

The Federal Aviation Administration (FAA) is slowly introducing rules for the use of drones in civilian airspace, and discovering, not unsurprisingly, that they cannot make everyone happy. When the FAA announced preliminary rules where small drones could only be flown within eyesight of the operator, Amazon immediately complained that this would ground their plans for one-day order delivery by drone. Soon after, the FAA recanted and allowed Amazon to test its drone delivery system. Unfortunately, our society has failed to have a full discussion as to which of the benefits that drones offer justify the risks drones present. Does the use of drones to facilitate police surveillance, to conduct research, for search and rescue operations, and to enable one-day order delivery justify crowded skies, air accidents, and continuing loss of privacy to swarms of camera-bearing mechanical insects buzzing overhead?

Finally, and most importantly, the scientists working in the field of artificial intelligence are waking up to the prospect of creating potentially dangerous artifacts that cannot be fully controlled. For decades, researchers in artificial intelligence had been bedeviled in getting their creations to perform basic tasks. Researchers who had been working in the field for 20 or 30 years, such as Cornell University’s Bart Selman, began to feel that “various forms of perceptions, such as computer vision and speech recognition” were “essentially unsolved (and possibility unsolvable) problems.”[1] In February 2009, Selman and Eric Horvitz, then President of the Association for the Advancement of Artificial Intelligence (AAAI), co-chaired the AAAI Presidential Panel on Long-Term AI Futures. At that gathering more speculative concerns, such as the possibility of smarter-than-human artificial intelligence, were explicitly addressed. "There was overall skepticism about the prospect of an intelligence explosion as well as of a 'coming singularity,' and also about the large-scale loss of control of intelligent systems," states the August 2009 interim report that summarized the meeting's conclusions. However, at a similar workshop in January 2015, Bart Selman stated, "a majority of AI researchers now express some concern about superintelligence due to recent breakthroughs in computer perception and learning."

The January 2015 workshop was organized by the *Future of Life Institute* to address growing concerns that, however distant, superintelligent artificial intelligence was possible, and to convey to scientists that it behooves them to begin work on approaches to ensure artificial intelligence systems would be demonstrably beneficial and controllable. Safe artificial intelligence suddenly became a concern for all artificial intelligence researchers, not merely those who had been focused upon narrow engineering tasks, machine ethics, robot ethics, or friendly artificial intelligence.

Research in a new approach to machine learning known as deep learning played an important role in catalyzing this uptick in concern. When DeepMind, a UK-based deep learning company, was acquired in 2014 by Google, its three founders made the establishment of an ethics board a condition for completing the sale. They recognized that they were unleashing a powerful technology that, while not dangerous presently, should never be appropriated for harmful purposes.

Deep learning utilizes neural networks, trained on massive computer systems, processing tremendous amounts of information. Such systems have demonstrated the capacity for learning complex rules without the human supervision and tweaking required by previous machine learning methods, including the ability to observe the data flow from other computers playing old video games such as *Breakout*. An algorithm designed at DeepMind developed its own strategies to play and creatively win a number of different computer games.

The primary focus for most of the articles reproduced in this volume are the practical ethical and legal considerations arising from the development of robots over the next twenty years. However, to properly frame your appreciation for the context in which those issues arises, we have elected to first add the most recent article. That is the statement on "Research Priorities for Robust and Beneficial Artificial Intelligence," which emerged out of the January, 2015 meeting organized by the *Future of Life Institute.*

Note

[1] Conveyed in a 2015 email to Wendell Wallach.

References not reproduced in this volume

Asaro, P. (2011). "A Body to Kick, But Still No Soul to Damn: Legal Perspectives on Robotics," in P. Lin, K. Abney, & G. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press, pp. 169–186.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford, UK: Oxford University Press.

Dennett, D. C. (1996, reprinted from 1991 version). "When Hal Kills, Who's to Blame?" in D. Stork (Ed.), *Hal's Legacy*. Cambridge, MA: MIT Press.

Gates, B. (2007), "A Robot in Every Home," *Scientific American*, January.

Gips, J. (1991). "Towards the Ethical Robot," in K. G. Ford, C. & P. Hayes (Eds.), *Android Epistemology* (pp. 243–252). Cambridge, MA: MIT Press.

Horvitz, E. & Selman, B. (2009). Interim report from the panel chairs: AAAI Presidential Panel on Long-Term AI Futures. Retrieved from https://www.aaai.org/Organization/Panel/panel-notes.pdf

Keynes, J. M. (1930). Economic Possibilities for our Grandchildren. Available online at: www.econ.yale.edu/smith/econ116a/keynes1.pdf/

Kurzweil, R. (2005). *The Singularity is Near. When Humans Transcend Biology.* New York, NY: Penguin Group.

Lehman-Wilzig, S. (1981). "Frankenstein Unbound: Towards a Legal Definition of Artificial Intelligence," *Futures* (December), 442–457.

Marcus, C. (2012). "Moral Machines," *The New Yorker,* November 24.

Wallach, W. & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

Wallach, W., Franklin, S. & Allen, C. (2010). "A Conceptual and Computational Model of Moral Decision Making in Humans and in AI," *TopiCS: Topics in Cognitive Science*, July.

White, L. F. & S. Chopra (2011). *Legal Theory for Autonomous Artificial Agents.* The University of Michigan Press.

Appendix 1

The Future of Life Institute: Research Priorities for Robust and Beneficial Artificial Intelligence: an Open Letter

http://futureoflife.org/misc/open_letter

Artificial intelligence (AI) research has explored a variety of problems and approaches since its inception, but for the last 20 years or so has been focused on the problems surrounding the construction of intelligent agents—systems that perceive and act in some environment. In this context, "intelligence" is related to statistical and economic notions of rationality—colloquially, the ability to make good decisions, plans, or inferences. The adoption of probabilistic and decision-theoretic representations and statistical learning methods has led to a large degree of integration and cross-fertilization among AI, machine learning, statistics, control theory, neuroscience, and other fields. The establishment of shared theoretical frameworks, combined with the availability of data and processing power, has yielded remarkable successes in various component tasks such as speech recognition, image classification, autonomous vehicles, machine translation, legged locomotion, and question–answering systems.

As capabilities in these areas and others cross the threshold from laboratory research to economically valuable technologies, a virtuous cycle takes hold whereby even small improvements in performance are worth large sums of money, prompting greater investments in research. There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. The potential benefits are huge, since everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty is not unfathomable. Because of the great potential of AI, it is important to research how to reap its benefits while avoiding potential pitfalls.

The progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximizing the societal benefit of AI. Such considerations motivated the AAAI 2008–09 Presidential Panel on Long-Term AI Futures and other projects on AI impacts, and constitute a significant expansion of the field of AI itself, which up to now has focused largely on techniques that are neutral with respect to purpose. We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial: our AI systems must do what we want them to do. The attached research priorities document gives many examples of such research directions that can help maximize the societal benefit of AI. This research is by necessity interdisciplinary, because it involves both society and AI. It ranges from economics, law, and philosophy to computer security, formal methods and, of course, various branches of AI itself.

In summary, we believe that research on how to make AI systems robust and beneficial is both important and timely, and that there are concrete research directions that can be pursued today.