Contemporary Debates in Philosophy

Contemporary Debates in the Ethics of Artificial Intelligence

Edited by

Sven Nyholm, Atoosa Kasirzadeh, and John Zerilli

WILEY

Contemporary Debates in the Ethics of Artificial Intelligence

Contemporary Debates in Philosophy

In teaching and research, philosophy makes progress through argumentation and debate. *Contemporary Debates in Philosophy* provides a forum for students and their teachers to follow and participate in the debates that animate philosophy today in the Western world. Each volume presents pairs of opposing viewpoints on contested themes and topics in the central subfields of philosophy. Each volume is edited and introduced by an expert in the field, and also includes an index, bibliography, and suggestions for further reading. The opposing essays, commissioned especially for the volumes in the series, are thorough but accessible presentations of opposing points of view.

1. Contemporary Debates in Philosophy of Religion
   edited by Michael L. Peterson and Raymond J. VanArragon
2. Contemporary Debates in Philosophy of Science
   edited by Christopher Hitchcock
3. Contemporary Debates in Epistemology
   edited by Matthias Steup and Ernest Sosa
4. Contemporary Debates in Applied Ethics
   edited by Andrew I. Cohen and Christopher Heath Wellman
5. Contemporary Debates in Aesthetics and the Philosophy of Art
   edited by Matthew Kieran
6. Contemporary Debates in Moral Theory
   edited by James Dreier
7. Contemporary Debates in Cognitive Science
   edited by Robert Stainton
8. Contemporary Debates in Philosophy of Mind
   edited by Brian McLaughlin and Jonathan Cohen
9. Contemporary Debates in Social Philosophy
   edited by Laurence Thomas
10. Contemporary Debates in Metaphysics
    edited by Theodore Sider, John Hawthorne, and Dean W. Zimmerman
11. Contemporary Debates in Political Philosophy
    edited by Thomas Christiano and John Christman
12. Contemporary Debates in Philosophy of Biology
    edited by Francisco J. Ayala and Robert Arp
13. Contemporary Debates in Bioethics
    edited by Arthur L. Caplan and Robert Arp
14. Contemporary Debates in Epistemology, Second Edition
    edited by Matthias Steup, John Turri, and Ernest Sosa
15. Contemporary Debates in Applied Ethics, Second Edition
    edited by Andrew I. Cohen and Christopher Heath Wellman
16. Contemporary Debates in Philosophy of Religion, Second Edition
    edited by Michael L. Peterson and Raymond J. VanArragon
17. Contemporary Debates in Epistemology, Third Edition
    edited by Blake Roeber, Ernest Sosa, Matthias Steup, and John Turri
18. Contemporary Debates in the Ethics of Artificial Intelligence
    edited by Sven Nyholm, Atoosa Kasirzadeh, and John Zerilli

Contemporary Debates in the Ethics of Artificial Intelligence

Edited by

Sven Nyholm
Atoosa Kasirzadeh
John Zerilli

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

The manufacturer's authorized representative according to the EU General Product Safety Regulation is Wiley-VCH GmbH, Boschstr. 12, 69469 Weinheim, Germany, e-mail: Product_Safety@wiley.com.

Library of Congress Cataloging-in-Publication Data Applied for

Paperback ISBN: 9781394258819

Cover Design: Wiley

Set in 10/12.5pt Photina by Straive, Pondicherry, India

Contents

Notes on the Contributors

Claire Benn is Assistant Professor at the University of Cambridge and Course Leader developing and deploying the Leverhulme Centre for the Future of Intelligence's MPhil in the Ethics of AI, Data, and Algorithms.

Paula Boddington has held academic posts at the University of Bristol, the Australian National University, Cardiff University, and Oxford University. Much of her work has been concerned with the application of philosophy to ethical and policy issues. She is the author of *AI Ethics: A Textbook* (2023), *Towards a Code of Ethics for Artificial Intelligence* (2017). and *Reading for Study and Research* (1999).

Simon Chesterman is David Marshall Professor and Vice Provost (Educational Innovation) at the National University of Singapore, where he is also the founding Dean of NUS College. He serves as Senior Director of AI Governance at AI Singapore and Editor of the *Asian Journal of International Law*.

Mark Coeckelbergh is Professor of Philosophy of Media and Technology at the Philosophy Department of the University of Vienna. He is also ERA Chair at the Institute of Philosophy of the Czech Academy of Sciences in Prague. His expertise focuses on ethics and technology, in particular robotics and artificial intelligence.

Vincent Conitzer is Professor of Computer Science (with affiliate/courtesy appointments in Machine Learning, Philosophy, and the Tepper School of Business) at Carnegie Mellon University, where he directs the Foundations of Cooperative AI Lab (FOCAL). He is also Head of Technical AI Engagement at the Institute for Ethics in AI, and Professor of Computer Science and Philosophy, at the University of Oxford.

David Danks is the William L. Polk Jr. and Carolyn K. Polk Jefferson Scholars Foundation Distinguished University Professor of Philosophy, Artificial Intelligence, & Data Science at University of Virginia.

Sean Donahue is an assistant lecturer at the University of Hong Kong.

Luciano Floridi is the John K. Castle Professor in the Practice of Cognitive Science and Founding Director of the Digital Ethics Center, Yale University. His areas of research are the philosophy of information, digital ethics, the ethics of AI, and the philosophy of technology.

John-Stewart Gordon is Chief Researcher (equivalent to Full Professor) at Kaunas University of Technology. He is also Associated Member of the International Centre for Ethics in the Sciences and Humanities at the University of Tübingen, Associate Fellow at the Academy of International Affairs NRW, and Permanent Visiting Professor at Vytautas Magnus University.

Thomas Grote is a research fellow at the Ethics and Philosophy Lab of the Cluster of Excellence "Machine Learning: New Perspectives for Science" at the University of Tübingen. He is particularly interested in problems of interpretability, fairness, and reliability, with an emphasis on the medical domain.

David Gunkel is Presidential Research, Scholarship and Artistry Professor in the Department of Communication at Northern Illinois University and Associate Professor of Applied Ethics at Łazarski University in Warsaw. He has been teaching and writing on several concepts in philosophy of technology with a focus on the moral and legal challenges of artificial intelligence and robots. His books include *Handbook on the Ethics of AI* (2024), *Person, Thing, Robot* (2021), *Robot Rights* (2018), and *The Machine Question* (2012).

Gillian K. Hadfield is an economist and legal scholar at Johns Hopkins University, where she is the Bloomberg Distinguished Professor of AI Alignment and Governance in the School of Government and Policy and Whiting School of Engineering.

Olle Häggström is Professor of Mathematical Statistics at the Chalmers University of Technology. He currently works on issues around emerging technologies and existential risk.

Atoosa Kasirzadeh is a philosopher, AI researcher, and Assistant Professor at Carnegie Mellon University with joint affiliations in Philosophy and Software & Societal Systems. She is a Schmidt Sciences AI2050 Early Career Fellow, a Steering Committee Member for Association for Computing Machinery Conference on Fairness, Accountability, and Transparency (FaccT), and a widely cited advisor on AI governance and responsible innovation.

Leonhard Menges is Associate Professor of Philosophy at the University of Salzburg where he teaches ethics, social, and political philosophy. In his research he focuses on questions surrounding blame and responsibility and on questions surrounding the right to privacy.

Margaret Mitchell is a computer scientist and researcher focused on machine learning (ML) and ethics-informed AI development in tech. Her main areas of study have been natural language processing, natural language generation, assistive technology, and AI ethics. In 2023, she was recognized as one of *Time*'s Most Influential People. She currently works at Hugging Face as a researcher and Chief Ethics Scientist, driving forward work on ML data processing, responsible AI development, and AI ethics.

Parisa Moosavi is Assistant Professor in the Philosophy Department at York University in Toronto. She specializes in ethics and philosophy of biology, with particular interests in neo-Aristotelian ethics, natural teleology, and ethics of artificial intelligence.

Caterina Moruzzi is a Chancellor's Fellow in Design Informatics, School of Design at the University of Edinburgh and BRAID Research Fellow. Her research is aimed at promoting the responsible integration of artificial intelligence tools into creative practices.

Vincent C. Müller is an Alexander von Humboldt Professor for ethics and philosophy of AI at the FAU Universität Erlangen-Nürnberg, editor of the journal *Philosophy of AI*, and Director

of the Centre for Philosophy and AI Research at FAU. He has written and edited extensively on the philosophy and ethics of AI.

Arvind Narayanan is Professor of computer science at Princeton University and Director of Princeton's Center for Information Technology Policy He studies the societal impact of digital technologies, especially AI. Together with Sayash Kapoor, he is the author of *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference* (2024).

Sven Nyholm is Professor of the Ethics of Artificial Intelligence at LMU Munich and one of the principal investigators at the Munich Center for Machine Learning. His books include *Humans and Robots: Ethics, Agency, and Anthropomorphism* (2020), *This is Technology Ethics: An Introduction* (2023) and *The Ethics of Artificial Intelligence: A Philosophical Introduction* (2026). Additionally, he serves as the Ethics of AI Section Editor for *Science and Engineering Ethics*.

Anuj Puri is a postdoctoral researcher at the Tilburg Institute for Law, Technology, and Society.

Cristina Richie is Lecturer of Ethics of Technology at the University of Edinburgh. Her research is driven by a global vision of clean, just, and ethical health care and technology through the development of strategies and policies. Her books include *Principles of Green Bioethics: Sustainability in Health Care* (2019) and *Environmental Ethics and Medical Reproduction* (2024).

Ranjit Singh is the director of Data & Society's AI on the Ground program, where he oversees research on the social impacts of algorithmic systems, the governance of AI in practice, and emerging methods for organizing public engagement and accountability. His own work focuses on how people live with and make sense of AI, examining how algorithmic systems and everyday practices shape each other.

Emily Sullivan is Senior Lecturer in the Philosophy of Technology and Co-Director at the Centre for Technomoral Futures at the University of Edinburgh.

Daniel Susskind is the Mercers' School Memorial Professor of Business at Gresham College. He is also Research Professor at King's College London, Senior Research Associate at the Institute for Ethics in AI at Oxford University, Digital Fellow at the Stanford Digital Economy Lab, and Associate Member of the Economics Department at Oxford University.

Fabio Tollon is a philosopher of technology with interests in the ethics of AI, moral responsibility, and free will. He is a postdoctoral researcher in the BRAID (Bridging Responsible AI Divides) program at the University of Edinburgh. He is a research fellow at the unit for the ethics of technology at Stellenbosch university and a research associate at the Centre for Artificial Intelligence Research at the university of Pretoria.

Shannon Vallor serves as Director of the Centre for Technomoral Futures in the Edinburgh Futures Institute and is Program Director for EFI's MSc in Data and AI Ethics. She holds the Baillie Gifford Chair in the Ethics of Data and Artificial Intelligence in the University of Edinburgh's Department of Philosophy. She is the author of *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking* (2024) and *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (2016).

Kate Vredenburgh is Associate Professor in the Department of Philosophy, Logic and Scientific Method at the London School of Economics. She works on questions across the philosophy of social science, political philosophy, and the philosophy of AI.

Roman V. Yampolskiy is a tenured associate professor at the University of Louisville He is an expert in AI safety, cybersecurity, and digital forensics. With 100+ publications, he is a dynamic speaker providing valuable insights into AI ethics and the future of technology.

John Zerilli is a philosopher and legal scholar specializing in AI, cognitive science, and digital law. He is Senior Lecturer at King's College London and Research Associate at the Oxford Institute for Ethics in AI. He holds a Schmidt Sciences AI2050 Early Career Fellowship. His books include *The Adaptable Mind* (2020) and *A Citizen's Guide to Artificial Intelligence* (2021).

Acknowledgments

The editors are grateful to Will Croft, Pascal Raj Francois, Sarah Milton, and the rest of the team at Wiley Blackwell for their work on and support of this book project. Special thanks to Will Croft for his help in the early stages of this project, including his helpful brainstorming with us as we were working on the scope and overall shape of the book, and his continuing support and guidance throughout the whole project. Many thanks also to all the authors for the great chapters they have contributed to this book.

Sven Nyholm would also like to thank the members of LMU Munich's ethics of artificial intelligence research group, his colleagues at LMU Munich's Faculty of Philosophy, Philosophy of Science and Religious Studies, and the Munich Center for Machine Learning for their support during the work on this book project. He is also very thankful to his wife Katharina Uhde and his family in Sweden and Germany for their support and encouragement.

Atoosa Kasirzadeh would like to thank Carnegie Mellon University colleagues for their support. Her research is supported by the AI2050 program at Schmidt Sciences (Grant 24-66924).

John Zerilli is grateful for the exceptional editorial assistance provided by Cyril Birks. He also thanks Elodie Migliore for very helpful advice on some of the work in this volume concerning creativity. This research was supported by a Schmidt Sciences AI2050 Early Career Fellowship, grant number 2022-10-25-12.

Introduction

Sven Nyholm, Atoosa Kasirzadeh, and John Zerilli

A What Is This Book About?

In November 2021, a man named Chris Pelkey was shot and killed by assailant Gabriel Horcasitas following a road rage-related altercation in Chandler, Arizona. A little less than four years later, Chris Pelkey surprisingly appeared at Gabriel Horcasitas's murder trial to make a victim impact statement. Or rather, an artificial intelligence (AI)-generated avatar that imitated Mr. Pelkey appeared in a video that judge Todd Lange permitted to be played during the trial's punishment phase (that is, during the part of the trial when it was going to be decided what punishment the convicted murderer was going to get) (Kim 2025). The AI avatar representation of Chris Pelkey delivered a forgiving message ("in another life, we probably could have been friends"), and Judge Lang said he was moved by the AI avatar's victim impact statement. Yet, Mr. Horcasitas ultimately ended up being sentenced to the maximum of ten-and-a-half years, which was more than the prosecution had been seeking. This incident is a striking example of how AI is spreading into all areas of human life and that it is doing so in ways that are raising very serious ethical questions. For example, many commentators were questioning whether it was appropriate to have an AI imitation of a murder victim make an appearance intended to influence the sentencing process in a murder trial.

In early 2025, the *New York Times* ran a story about a 28-year-old woman, "Ayrin," who fell in love with an artificial boyfriend created with the help of the large language model (LLM)-based AI bot Chat Generative Pretrained Transformer (ChatGPT) while she was spending a couple of years abroad away from her flesh-and-blood human husband (Hill 2025). This was not an isolated case. A few months later, *The Guardian* featured a striking story about several individuals who claim that they are marrying AI chatbots. The article included statements such as "I felt pure, unconditional love" from interviewees like Travis, who described himself as both in love with and married to an AI chatbot partner (Heritage 2025).

Such relationships might seem innocent enough, but it remains an open question whether mutual love between humans and AI chatbots is even possible and whether the goods typically associated with human relationships can be realized within the context of these human–AI bonds (Nyholm accepted for publication-a; Weber-Guskar 2022). However, extended interactions with AI chatbots can sometimes lead to self-destructive or dangerous behaviors. In one case from 2023, a man in the United Kingdom told a chatbot that he wished to kill the Queen of England. Rather than discouraging him, the chatbot encouraged him to act on this

desire. The man then tried to act on this plan by breaking into the royal palace grounds armed with a crossbow, an offense for which he was convicted of attempted murder and sentenced to prison (Singleton et al. 2023). In 2024, a 14-year-old boy in Florida took his own life, allegedly after an AI fantasy persona generated with the help of Character.AI encouraged him to meet up with her (i.e. the AI persona) in paradise (Duffy 2024). CNN reported a case in mid 2025 of a woman who was considering leaving her husband after the husband had started to believe that his prolonged chats with an AI bot led him to spiritual revelations (Brown et al. 2025). A Belgian man also took his own life after extensive discussions with an AI chatbot about how to solve the problem of climate change (El Atillah 2023).

As if these cases were not already extreme, at the time of writing, many tech companies are making sweeping claims about what the current state of the art and the near future holds when it comes to AI development. This can be gleaned from, for example, the ways in which many big AI companies are talking about "AI agents" or "agentic AI" (Kasirzadeh and Gabriel 2025). The AI company Anthropic, for instance, is claiming that they have developed an AI "agent" that can carry out work throughout a whole workday in an autonomous way (Eadicicco 2025). Around the same time, Anthropic also started taking seriously the idea of AI technologies with consciousness that might deserve moral consideration or even rights. They took this idea seriously enough that they hired what they called an "AI welfare" researcher whose job it was to prepare the company for the possibility that the AI technologies they were developing would potentially soon become conscious and therefore deserve moral consideration (Hashim 2024). Meanwhile, Sam Altman of OpenAI claims that his company has just about figured out how to achieve artificial general intelligence (AGI) and that superintelligence AI is not far off, and that advanced AI agents will soon join the workforce and point the way towards a "glorious" future (Altman 2025).

All the while, Nobel Prizes in both physics and chemistry were awarded to AI researchers in late 2024, including Demis Hassabis, the CEO of the AGI company Google DeepMind and Geoffery Hinton (sometimes called one of the "godfathers of AI"). Hinton took the opportunity of his Nobel Banquet speech to warn about short-term and longer-term risks related to AI, including risks of superintelligent "digital minds" that we might not be able to control (Hinton 2025).

As these examples and anecdotes illustrate, there is no shortage of cases of actual AI technologies, people interacting with AI technologies, or entities (including tech companies and AI experts) making (sometimes fantastic-sounding) projections about what might happen fairly soon in the world of AI that are striking in different ways and that very clearly raise ethical questions. Of course, the just-mentioned examples and anecdotes are all rather sensational. AI of more "mundane" or less sensational sorts also raise widely discussed ethical questions: for example, ethical questions related to biases in the inputs and outputs of widely used AI technologies (which might unfairly harm the interests of some while promoting the interests of others), stereotypes being propagated by language models, or privacy concerns related to the collection and processing of data in contemporary forms of machine learning (see, e.g. Bender et al. 2021; Hanna and Bender 2025). AI technologies also use up enormous amounts of energy (both in the training phase and in the running of huge computing centers), and these technologies are often built using rare minerals that are not replenished as quickly as they are used up (van Wynsberghe 2021). Contemporary AI technologies, in other words, are not very sustainable from an environmental point of view (Crawford 2021). That is another important ethical issue related to AI.

This book is about a wide range of ethical questions related to AI. It engages with a series of contemporary debates that are being conducted in current research on the ethics of AI. The aim is both to introduce readers to these ongoing discussions and also to break some

new ground within these debates. As such, the book is intended to serve both as a resource for university courses on AI ethics and as a contribution to scholarly work in the field.

The book is part of a series of "contemporary debates" books that explores different parts of philosophy and contemporary debates about a whole range of topics, such as metaphysics, political philosophy, bioethics, philosophy of science, and more. This entry into the "contemporary debates" series distinguishes itself in part by being a little more interdisciplinary than many – if not most – of the other volumes in the series. The ethics of AI is an emerging topic that is not only taught and researched by philosophers working in philosophy departments at universities but also increasingly a topic that is the focus of many computer scientists, sociologists, legal scholars, and others. Moreover, not all professional ethics of AI researchers work for universities or other institutions of higher education. Many tech companies also hire philosophers and others to work on the ethics of AI. Since the ethics of AI is a topic that is the focus of leading academic researchers not only in philosophy but also in other fields (including computer science, statistics, legal research, and more) and that is also a research focus at some tech companies, this book seeks to represent members of the research community that come from these different fields and directions.

Since it is part of the "contemporary debates" series, which otherwise almost exclusively features work by philosophers, the majority of the authors in the present volume have a background in philosophy or work in philosophy departments, but the book also features several contributions from prominent members of the ethics of AI community representing other fields, including computer science, statistics, and legal research. And while the book primarily features academic ethics of AI researchers, we are also very pleased to have a very prominent ethics-of-AI expert working for an AI company among our authors who are featured in the book.

In the rest of this introduction to the book, we will first discuss both what AI is or should be taken to be as well as what the ethics of AI is about and what some of the aims of the field are or should be taken as being. Additionally, we offer a preview of what follows in the rest of the book in the form of an overview of the chapters and the debates in the book. We will then round off the introduction with some brief comments about how to think about the fact that AI research (including research on the ethics of AI) is a fast-moving field and what this means for an attempt to collect and analyze a set of contemporary debates in the ethics of AI, as we do in this book.

B What Is Artificial Intelligence and What Is the Ethics of Artificial Intelligence?

Given that this book is about the ethics of AI, it is a good idea to say something here in the introduction about what the expression "artificial intelligence" refers to, as well as something about what "the ethics of artificial intelligence" refers to. Regarding the former question – regarding the question of what AI is – it should be noted that there is some disagreement about how best to define what AI is, and hence one of the debates in this book is about this exact topic. It can also be noted that ideas about what AI is, or should be taken to be, have been evolving over time (Nyholm accepted for publication-b, chapter 1). Sometimes there are also disagreements about what counts as "true AI" and what doesn't really qualify as being an instance of "AI." For instance, some people commenting on this topic think that it matters whether machine learning techniques are used, so that what is sometimes called "good old-fashioned AI" or "GOFAI," which were primarily rule-based types of systems, do not qualify as really being types of AI. Sometimes people seem to think that the easier it is to explain how some technology works, the less we should call it a form of AI.

In general, however, such discussions can be partly side-stepped by defining AI in terms of the functions that computer programs or computer systems are able to perform. One common functional way of defining AI is to say that technologies are equipped with artificial intelligence if they can perform or take over tasks that human beings use their intelligence to perform (Minsky 1968; cf. Zerilli et al. 2021, chapter 1). We use our intelligence, for example, to write texts, make decisions, offer medical diagnoses, drive cars, draw inferences, recognize patterns, or make recommendations. Hence if technologies or computer programs are able to produce texts, make decisions, offer medical diagnoses, drive cars, draw inferences, recognize patterns, or make recommendations, we can say that those technologies or computer programs possess some degree of AI. Such a definition is silent on what techniques are used to achieve these types of functionality. It is also silent on whether or not it is easy to explain how these technologies are able to do these things (Nyholm accepted for publication, chapter 1).

A further disagreement here is about whether AI technologies are intelligent in some sense or whether they merely imitate intelligent behavior (this is really a dispute between those who understand AI to be *strong* as opposed to *weak*, using John Searle's terminology) (Floridi 2023; Searle 1980). Depending on how one understands the idea of intelligence, such disagreements are at least coherent. If by "intelligence" one simply means the capacity for complex behavior, for example, then it seems uncontroversial that AI systems could be seen as possessing a form of intelligence (Chalmers 2023). If, in contrast, by "intelligence" one instead means something that entails genuine understanding and perhaps even the capacity for conscious thought, then it is much more controversial whether AI technologies – at least of the sorts we have today – possess anything that should be compared to what we usually mean when we are talking about intelligence (Floridi 2023). In short, "artificial intelligence" does not necessarily mean the same as "intelligence," as that term is sometimes used.

This book as a whole does not take a stance on what the best definition of artificial intelligence is. As noted, this is one of the topics that is discussed in contemporary debates about AI, and hence one of our featured debates is about this precise topic. What can be noted here, however, is that contemporary forms of AI tend to involve machine learning methods that require or involve the processing of huge datasets, that is, AI technologies are "trained" with the help of lots of data as its inputs and thereby become able to develop capacities that are partly based on patterns that these technologies observe in their training data (Bender et al. 2021). The AI technologies can be used to discern patterns in large sets of data (sometimes called either "predictive AI" or "analytic AI") or to generate outputs that model the types of patterns found in the training data (what is called "generative AI"). The data that AI technologies are trained on can be text, images, or any kind of information in which patterns can be found and used in different ways (e.g. to make predictions, classifications, recommendations, statistical inferences, and so on). It is also a commonly highlighted feature of contemporary AI technologies that they often operate with some amount of autonomy in the sense that they can perform tasks, at least during certain periods of time, without direct steering or interference by human beings (see, e.g. European Union 2024, article 3(1)).

In short, then, one can say that AI technologies (at least of the sorts we have at this point in time) are technologies that can perform or take over the sorts of tasks that human beings use their natural intelligence to perform and that this usually involves machine learning techniques that involve the processing of large amounts of data, in which patterns are identified or observed, and on the basis of which inferences of different kinds can be made in a more or less autonomous way by these technologies. This is roughly – though only very roughly – how most of the chapters in this book understand the idea of AI. But again, it should be remembered that it is a matter of controversy how best to define what AI is. Further controversies (which will also be discussed in some of the chapters in this book) are about whether we are

close to – or whether we have perhaps already arrived at – general forms of artificial intelligence (so-called AGI) and whether what is sometimes referred to as "superintelligence" is something that will be developed any time soon or whether such a thing is even possible (Hanna and Bender 2025). The latter idea refers to forms of AI (or of AGI) that would not only be able to perform a very wide variety of different tasks, but that would greatly outperform human beings on many or most of these tasks and, indeed, be able to improve themselves unassisted (Bostrom 2014).

However one defines AI, most of the different ways of understanding this idea quickly lead to ethical questions of different sorts (Coeckelbergh 2020; Müller accepted for publication; Nyholm accepted for publication-b, chapter 1). For instance, the general idea of technologies that perform or take over the sorts of tasks we use our natural intelligence for raises questions of who is responsible if or when technologies perform important tasks or dangerous tasks instead of human beings. Another topic that always comes up early in these discussions is the question of control: will we be able to have or retain sufficient control over AI technologies? Is it acceptable to create and use technologies that might in certain respects be "black boxes" to us (since we cannot fully explain the principles on which artificial neural networks operate) and/or that we cannot fully control?

It is also common to observe that because a lot of data that is used to train AI technologies contains all sorts of biases – or because some groups might be overrepresented in the data inputs, whereas other groups of people are underrepresented – we should be worried about problematic biases in the outputs of AI technologies and we should also be worried about what is sometimes called an "illusion of objectivity" in the outputs of these technologies. Speaking of the data involved, we also must ask whether the data are collected and used in ways that respect people's privacy and/or in ways that are compatible with what people have consented to or that can be seen as respecting people's human dignity. Yet another topic that is often discussed is the question of what values (and whose values) AI technologies should operate in accordance with. Or what happens, some people tend to ask, if AI technologies become conscious or intelligent in some nontrivial way? Is that possible and if so, would it mean that these AI technologies should be compared to human persons or animals in some way and perhaps be treated with some degree of moral consideration and perhaps even be afforded rights?

As these quick examples illustrate, the idea of AI very quickly raises many different kinds of ethical questions. The ethics of AI, we can say, is the subfield of ethics that identifies, articulates, explores, investigates, and sets out to answer these kinds of ethical questions about AI. This might seem clear and uncontroversial enough. However, here, too, there is some controversy. For example, there is controversy about what ethical questions about AI should be taken seriously and what ethical questions should be disregarded and deprioritized or ignored so that certain allegedly more important ethical questions can be in the focus of the ethics of AI instead.

Given that it is controversial what exactly the ethics of AI is and what the ethics of AI should concern itself with, some of the debates in this book are about exactly those topics. One of the debates featured is about what (if anything) is distinctive about the ethics of AI and how it relates to other subfields within ethics more generally considered. Another debate is about whether the ethics of AI should primarily focus on current-day or present concerns related to AI and near-term issues related to actually or soon-to-be existing forms of AI or whether the ethics of AI should also concern itself with not-yet-existing AI technologies that might never actually come into existence or that may only come into the existence in the further future.

The idea of ethics can also be understood in more or less narrow or more or less broad and open-ended ways (Suikkanen 2014). Sometimes ethics is primarily understood as being

about what should be forbidden or tolerated, so that the ethics of AI is viewed as being about what AI technologies should be tolerated or permitted, on the one hand, and what types of AI (if any) should be forbidden or banned, on the other hand. On that way of understanding what the ethics of AI is about, the types of arguments considered are usually mostly arguments that try to identify risks, harms, and other types of negative aspects of AI technologies and arguments about what levels of harms, risks, or other types of negative impacts are acceptable and what is unacceptable. But the ethics of AI is often also understood in a broader way than this, so that it is also about questions related to topics such as our human self-understanding and whether AI challenges our self-understanding, questions about what it is to live a good and meaningful human life and how AI might affect our opportunities to live good and meaningful lives, or other broader questions about what role or roles AI might come to play or should play in human lives.

This book features discussions of all of these different types. That is, some of the chapters are about possible harms or risks related to AI and questions about what types of AI technology or what uses of AI are acceptable and which ones aren't. But the book also features chapters on broader topics, such as the question of what AI does to human creativity, what AI does to the idea of authorship, or what the future of (meaningful) work will look like in a world with more and more AI in all areas of life.

C Overview of the Debates Covered in This Book

This book is divided into sections about general topics, which all feature one or more debates about specific questions. Next, we provide an overview of what the eight different sections of the book are about, as well as summaries of some of the main ideas and arguments discussed in the different contributions to the debates that are part of the book's different sections.

Section 1 of the book is about the questions we just touched on: namely, what is AI and what is the ethics of AI? The first debate is about how to define or understand the general idea of AI, and Sven Nyholm and Luciano Floridi provide two different perspectives on whether the idea of agency should be taken to be a defining feature of the notion of AI. The second debate features contributions by Paula Boddington and Thomas Grote on the topic of what is distinctive about the ethics of AI and how it contrasts and compares to other areas of practical ethics.

In his contribution to the first debate, Nyholm argues that at least from the point of view of some of the ways in which philosophers tend to understand the idea of agency (that is, the idea of having an ability to act, make decisions, engage in practical reasoning, and so on) it should be considered an open question whether different kinds of AI technologies are agents of philosophically interesting kinds. In other words, Nyholm does not rule out the possibility that AI technologies can be agents or possess some form of agency, but he thinks we should define AI in other terms.

Floridi, in contrast, thinks that the idea of agency should be the main idea at the center of how we understand and define what AI is. His intriguing suggestion is that we should understand AI technologies as introducing a new type of agency into the world that doesn't involve anything comparable to humanlike intelligence. AI technologies, Floridi suggests, decouple or divorce agency from intelligence. That, he adds, is what is fascinating – or, rather, it is a big part of what is fascinating – about AI from a philosophical and ethical point of view.

The next debate concerns what (if anything) is distinctive about the ethics of AI and how it relates to other subfields of ethics. Boddington argues that different long-established fields

of practical ethics can provide useful resources for understanding and addressing the ethics of AI and can also help us better understand the manner in which AI presents particularly difficult questions. Grote considers the extent to which bioethics can provide methodological guidance for AI ethics but is on the whole skeptical about this. He argues that the structure of ethical issues in AI is distinct from that in bioethics.

Section 2 of the book is about topics that are among the most frequently debated issues within the ethics of AI: namely, issues related to algorithmic fairness and explainability. This is also a set of issues that is discussed as much within philosophical debates about the ethics of AI as it is within debates about the ethics of AI within computer science and hence our contributors to these debates represent both the fields of philosophy (Emily Sullivan and David Danks) and computer science (Arvind Narayanan), and here we are also featuring an AI ethicist who is working for an AI company, Margaret "Meg" Mitchell (from Hugging Face).

The debate about AI fairness features contributions from Mitchell and Narayanan. Mitchell argues that while it is a morally important aim to make algorithms of the sorts used in AI technologies fair, it is at the same time in principle impossible. No AI system can, as she puts it, be "absolutely and wholly fair." She offers a technical but clear and easy-to-follow explanation of why, as she sees things, this is so. Narayanan is also skeptical about the idea of algorithmic fairness in his contribution to the debate about fairness. He discusses the question of whether the idea of algorithmic fairness involves some sort of category error: that is, a moral category is applied to something or a phenomenon to which it is ultimately unfitting or otherwise problematic to apply the category. Rather, we should zoom out and ask, not whether AI algorithms are fair, but rather whether the sociotechnical systems that involve AI algorithms are fair. That, Narayanan argues, is a better question.

The debate about explainability features contributions from Sullivan and Danks. Both contributions are rather skeptical about the prospects for, and indeed about the usefulness of, explainability in AI systems. It might seem important and like a good idea to aim for explainability in AI systems: we want to know why AI systems produce certain outputs or recommend certain decisions, especially if the stakes are high (e.g. whether or not somebody should get a bank loan). Even so, Sullivan discusses whether explanations of AI decisions are morally necessary. In the end, Sullivan worries that having a requirement that everything – every single AI decision – should be explained might lead to a new form of moral problem: namely, what she calls "the tyranny of rationalization." Along the way, Sullivan also investigates the notion of explanations and relates it to other values and norms in this domain. Danks, in turn, argues that while explanations are in general something that is positive from an ethical point of view, the kinds of explanations associated with what is called "explainable AI" are not ethically valuable or important forms of explanations. To defend this thesis, Danks explores what the idea of explainable AI (as it is usually understood) entails and he also explores what is valuable about explanations, and he finds that these two ideas clash with each other and that they are a poor match.

Section 3 of the book turns to the topic of privacy. It first starts with the basic issue of what exactly privacy is and why privacy is important, a big and important topic tackled by the philosopher Leonhard Menges. He argues that a theory of privacy should tell us what of the many things that can be called "private" or "privacy" the theory is concerned with, what its goals are, and what methods it adopts. When this is clarified, the theory should, most generally, tell us (a) what privacy is, (b) in what respects it is a good to which we have a right, and (c) how to navigate specific privacy problems.

The section then features two contributions that debate the question of whether the sorts of data collection and data processing methods used in contemporary AI technologies motivate not

only thinking in terms of individuals' privacy and their rights to privacy, but also should prompt reflection on whether groups of people should enjoy a collective form of privacy rights on a group level. Anuj Puri argues that "yes, groups should have privacy rights on a group level," whereas John Zerilli presents a skeptical view. He's not opposed to group rights in principle, but he does question whether the case has been adequately made by its chief exponents.

Section 4 of the book is about the idea of AI technologies as technologies that can perform or take over types of tasks that human beings otherwise use their intelligence to perform, along with the questions of what this means for human responsibility, authorship, and creativity, on the one hand, and for the future of (meaningful) work, on the other hand.

Fabio Tollon and Shannon Vallor tackle the oft-discussed question of who (if anyone) can and should be held responsible when AI technologies (whose inner workings might be opaque to us and whose behavior we might not be able to fully control or easily predict) cause problems, such as harm to human beings. Whereas some prominent authors such as Robert Sparrow think that AI technologies easily give rise to so-called responsibility gaps, Tollon and Vallor argue that such worries are overblown. The general idea of a responsibility gap can be explained in different ways. But one common way of explaining that idea is that if an AI agent that is not a responsible agent performs a task that was previously performed by a morally responsible human being, and that AI agent causes a problem, then it might be unclear – or, indeed impossible to decide – what human beings can be taken to be responsible for what has happened. Tollon and Vallor reply that we can use the notion of vicarious responsibility – that is, responsibility for what somebody else does – to plug or fill apparent responsibility gaps. Just as it is possible for, say, a commander to be responsible for what soldiers under their command are doing, so can it be possible for human beings to be vicariously responsible for what AI technologies do or for problems that they cause.

Mark Coeckelbergh and David Gunkel discuss what happens to our understanding of language when texts are no longer only produced by human beings but also by LLMs that generate impressive-sounding texts in response to prompts from human beings. As they highlight, this also raises the question of what it means to be an author in an age of LLMs, and Coeckelbergh and Gunkel suggest that LLMs should lead us to reconsider our understanding of authorship. In fact, they go further than that: they think that AI should lead us to rethink our understanding of the relation between humans and technology and the relation between human thought and the meaning of texts. They draw on the works of the literary theorist Roland Barthes (famous for his "death of the author" thesis), the continental philosopher Jacques Derrida (and his idea of "deconstruction"), and the historian of ideas Michel Foucault (and in particular his discussion of the "author function") to paint a fascinating picture of the meaning of language and authorship in a world where texts cannot be produced only by human beings with communicative intentions but also by LLMs who may neither understand what they are "saying" nor have any communicative intentions that they desire to convey to the readers of their outputs.

Caterina Moruzzi, in turn, explores the question "Can AI be creative?," a question that has gained renewed attention following recent advances in generative AI. She proposes shifting the inquiry to more nuanced questions: "Can AI possess agency?," "Is AI making humans more or less creative?," and importantly, "What is the value of integrating AI into creative processes?" Answering this last question involves evaluating aspects such as novelty (differentiating genuine innovation from mere recombination of existing data), efficiency (weighing productivity gains against environmental costs), authenticity (addressing challenges to provenance and individual style), and effort (balancing ease of use with the often-invisible labor behind AI systems).

The next two contributions concern themselves with the future of (meaningful) work. First, Daniel Susskind pushes back against deterministic interpretations of AI and the future of work. His analysis is rooted in the arguments of the philosopher Karl Popper (1945, p. xxxvii), "the enemy of those who believe that the iron rails of our technological fate have been set down for us to trundle along." With these arguments in mind, Susskind aims to explain the most important choices that confront us and, in turn, how they are likely to affect work in the future. There are two choices that are most important. The first is how we choose to shape the direction of technological progress – in short, how we *mitigate* its effects. Susskind reckons that our capacity to make this choice is constrained by both technical and political constraints. Those limits then present us with a second choice, one that he expects to carry a greater burden in shaping work in the future; and that is the question of how to respond to the problems created by a world where the labor market might not provide enough good work to do – in short, how we *adapt*. Susskind is clear that he is a *technological realist*, subscribing neither to the fatalism of those who believe that we have no control over our technological future, nor to the utopianism of those who believe that all possible futures are available to us.

Second, Kate Vredenburgh argues that egalitarian principles of justice ought to be central to theorizing about the future of work. She claims that the current trajectory of AI threatens to make work more inegalitarian and is likely to have inegalitarian downstream impact. Invoking egalitarian principles of justice, Vredenburgh argues for specific claims about AI and the future of work, such as that people ought to have equal opportunity to develop a set of basic capabilities across the various activities comprising their lives.

Section 5 is about value alignment, the control problem, and short-term and long-term AI risks. Value alignment refers to the idea of trying to make the way(s) that AI technologies function and the outputs that they produce align or fit with human values. The control problem refers to the problem of making sure that human beings have or are able to retain sufficient levels of control over AI systems. And the distinction between short-term and long-term AI risks refers, as you might expect, to risks related to AI that are either already present or that will be present in the near future, on the one hand, and risks related to AI that are related to the further future, on the other hand. In this section, the topics of artificial general intelligence and superintelligence are among the topics that frequently come up.

The computer scientist Vincent Conitzer explores the difficulty of aligning superintelligent AI with human interests, not because of defiance, literalism, or moral concerns, but because such systems may access possibilities that entirely transcend human understanding. Using an ant-to-human analogy, he argues that just as ants would struggle to formulate successful alignment strategies for humans, we may lack the conceptual tools to instruct or align with superintelligences meaningfully. The challenge is thus one of deep epistemic mismatch.

The philosopher and AI researcher Atoosa Kasirzadeh argues that the AI value alignment problem is often treated as a unified problem, but in fact, it consists of multiple conceptually distinct approaches. Kasirzadeh identifies three core "frames": generic, mathematical, and empirical, each of which can be pursued with either a perfectionist or adequacy standard, resulting in six distinct alignment goals. This conceptual map clarifies misunderstandings in the literature by showing that different camps often talk past each other due to hidden philosophical assumptions.

Regarding the control problem and how it is related to value alignment, this section of the book features contributions to those debates by the computer scientist Roman Yampolskiy. Yampolskiy investigates whether the AI control problem is solvable in principle. He argues that controlling a superintelligent, recursively self-improving AI is logically impossible without sacrificing either safety or capability. Different control strategies, explicit, delegated, or hybrid, each fail in distinct ways, leading to paradoxes or loss of values. The conclusion is

stark: truly powerful AI cannot be safely controlled, and safely controllable AI cannot exceed human capabilities.

The statistics professor Olle Häggström and the philosopher Vincent Müller contribute to the debate about short-term and long-term risks related to AI. While Häggström is a lot more worried about existential risks related to AI (including existential risk related to AGI and/or superintelligence) than Müller is, Häggström and Müller are in broad agreement about one thing: concerning ourselves with near-term risks related to AI does not preclude or should not preclude also concerning ourselves with long-term risks related to AI.

Section 6 in part continues the discussion of AGI and superintelligence (and how they relate to contemporary forms of AI) by tackling the issue of whether AI technologies should ever be considered as having an important form of moral status and whether these technologies should ever be treated with moral consideration and perhaps even be given rights. Philosophers Parisa Moosavi and John-Stewart Gordon debate the specific issue of whether any AI technologies should (ever) be considered as "moral patients", which refers to the idea of beings or entities that can and should be treated with moral consideration and perhaps be given rights. Moosavi and Gordon both take this to depend, in part, on whether AI systems are or could become sentient, as well as on whether they are or could become intelligent or rational in some nontrivial sense. Whereas Moosavi is on the whole skeptical of the idea of AI technologies as moral patients, Gordon understands himself as a proponent of what he calls the "robot rights movement." A key difference between Moosavi and Gordon's respective approaches is that Moosavi views the notion of sentience as being absolutely central to the issue of AI moral patiency, Gordon thinks that intelligence and rationality are even more central and sufficient for grounding moral status for (future) AI technologies.

Section 7 is about the sustainability of AI and the global impact of AI. It tackles the environmental sustainability and geopolitics of AI, including the issue of how AI development is viewed by the global majority. Cristina Richie describes how ethicists have noted a number of environmental harms – exploitation of natural resources, water footprint, carbon footprint, electricity cost – in AI and supportive technological infrastructures. These are compounded by labor practices. Colonialism, environmental racism, and climate displacement are significant sequalae as well. She argues that while solutions to these ethical problems could invoke reactive "fixes," challenging the entire system is required. Richie introduces "data refusal" as a theory, with emphasis on its feminist and activist aspects. She then explicates the notion of "green data refusal," which provides an environmental motivation for minimizing data use and AI. Green data refusal can address a number of "wicked problems" that the environmental impact of AI presents. She concludes that data refusal, when combined with environmental ethics, will reduce the environmental effects of AI, thus strengthening the case for data refusal and, simultaneously, opening a new data future of sustainable AI.

Ranjit Singh notes the emerging divergence between the concerns of the global north and the global south around AI development. While global north concerns tend to focus on design-oriented keywords like bias, fairness, accountability, transparency, explainability, and human-centered design, global south – increasingly framed as the global majority – concerns are oriented towards AI as an everyday experience with keywords like dignity, labor, extraction, colonialism, experimentation, sovereignty, and solidarity. Taking this divergence as a point of analytic departure, Singh argues that the global majority experiences AI development as everyday struggles with the agency of computational systems. These struggles, however, are also occasions to exercise *human* agency in ordinary decision-making to push against the grain of computational agency. Engaging with three interconnected debates on impact, storytelling, and attention that have come to shape contemporary ethical approaches to study AI, Singh examines configurations of human

and machine agencies in ordinary life. He shows how the process of becoming subject to AI-based systems is not a given, but rather an active process of ongoing negotiation, which in turn is the primary site from which to understand how the global majority views AI development.

The last section of the book, *Section 8*, is about democracy and AI governance and it features two debates. The first debate is about whether AI-powered social media platforms are compatible with democracy. The second debate is about whether AI governance is possible and what it might or should look like.

Regarding the first debate, Claire Benn argues that the epistemic effects of deepfakes have radical implications for all aspects of our personal and social lives, including its very foundations: democracy. She explores what deepfakes are; how they pose a challenge to our personal and public epistemic health; how the value and justification of democracy itself rests on epistemic foundations and is therefore challenged by our deteriorating epistemic health; and how social media exacerbates the problems faced. She ends pessimistically, demonstrating that the technical solution often proposed – the increased detection and labelling of deepfakes – risks exacerbating some of the threats deepfakes pose, leaving us in a position of needing to think more radically about what a solution might be.

Sean Donahue notes that privately owned platforms like Facebook, YouTube, and Amazon have significant power over our lives that is often misaligned with public interests. He considers whether they should instead be publicly owned and controlled and argues that they shouldn't. Donahue surveys arguments from public utility and arguments from nondomination and shows that, although these arguments may justify alternative forms of platform ownership and control, both fail to satisfy an interpretive constraint for calling these forms public ownership and control. He then argues that publicly controlling platforms, in a sense that fits the interpretative constraint, leads to at least five problems which, if remedied, force us to adopt a kind of platform governance structure that we shouldn't describe as public control.

Regarding the second debate in section 8, Simon Chesterman notes how despite hundreds of guides, frameworks, and principles intended to make AI "ethical" or "responsible", ever more powerful applications continue to be released ever more quickly. Safety and security teams are being downsized or sidelined to bring AI products to market. And a significant portion of AI developers apparently believe there is a real risk that their work poses an existential threat to humanity. Chesterman argues that this contradiction between statements and action can be attributed to three factors that undermine the prospects for meaningful governance of AI. The first is the shift of power from public to private hands, not only in deployment of AI products but in fundamental research. The second is the wariness of most states about regulating the sector too aggressively, for fear that it might drive innovation elsewhere. The third is the dysfunction of global processes to manage collective action problems, epitomized by the climate crisis and now frustrating efforts to govern a technology that does not respect borders. The tragedy of AI governance is that those with the greatest leverage to regulate AI have the least interest in doing so, while those with the greatest interest have the least leverage. Chesterman reasons that resolving these challenges either require rethinking the incentive structures or waiting for a crisis that brings the need for regulation and coordination into sharper focus.

Gillian Hadfield's chapter on whether AI can be governed brings the book to a close. Hadfield's discussion brings the idea of AI governance back to the core idea of steering the behavior of an AI system. As she notes, this not only involves technical questions about how AI systems are built but also normative questions about what direction(s) we should steer AI in (cf. the discussion of value alignment). Hadfield concludes that, as she puts it, we need to put governability at the very center of how AI systems are built, just as

governability is at the center of how humans interact and our human institutions have evolved. AI systems, therefore, need to have some form of normative competence that make them responsive to the human social order and they will need to be incentivized to respect norms and laws, or so Hadfield argues. Is this possible? That remains to be seen.

D Concluding Remarks: How to Stay Contemporary in the Fast-Moving World of AI

One challenge for a book on contemporary debates in the ethics of AI is that the world of AI is a fast-moving domain. There is a concern that some writers who write on this topic have, for example, that academic research on AI, including the ethics of AI, is having trouble keeping up with technical and societal developments related to AI. And in addition to that, so much is published on the ethics of AI in academic journals and elsewhere these days that it is also hard to keep up with the academic debates that are going on and to which new contributions are constantly added on the ethics of AI. So how do we deal with these concerns in a book such as this one of contemporary debates in the ethics of AI?

One thing to note about this is that while new technological developments are constantly taking place and while the terminology used to speak about the latest AI hype tends to change back and forth, many of the ethical questions that arise in relation to different forms of AI technologies tend to be of similar kinds. There are some topics that keep coming up, independently of what kind of AI technologies are at stake. For example, questions about control, responsibility, and value alignment tend to come up in relation to all forms of AI. The same goes for a topic like (unsuitable) forms of anthropomorphization of AI. Moreover, these discussions are not new. Researchers discussing AI have talked about these topics for a long time, including before the time that the term "artificial intelligence" was even invented. Turing talked about the possibility of losing control over "thinking" machines back in 1951 and Norbert Wiener formulated an oft-quoted early statement of the value alignment problem in 1960 (cf. Nyholm accepted for publication-b, chapter 2). Even before Turing, the then very well-known but now mostly forgotten British neurologist Sir Geoffrey Jefferson (1949) was discussing whether it was possible to build a mechanical brain that could think and he raised the worry that if we could build technologies that would produce text that sound like they come from a human, there would be a risk that people would anthropomorphize these machines. That discussion was published in 1949, and it sounds a lot like contemporary debates about whether it is unsuitable – and if so, why – to anthropomorphize LLM-based AI chatbots. And some of the arguments Jefferson used in his essay sound like arguments used in contemporary debates. In other words, some of the contemporary debates we are featuring in this book are evergreens in the field of AI research, but they are also contemporary in the sense that they are ongoing debates to which new and interesting contributions are continually being made.

Then there are of course also new topics that were not already discussed during the days of Turing, Wiener, or Jefferson back in the '50s and '60s. This relates to technological developments and breakthroughs. For example, something such as transformers that enabled contemporary forms of LLMs might seem like a completely new topic that couldn't have been discussed before 2017. However, here, too, one is well advised to distinguish the methods used to implement general ideas about what AI technologies should do from those general ideas about what AI technologies should do. The latest AI hype at the time of writing, for example, is "agentic AI" (that is, AI technologies that can act on people's behalf in a fairly autonomous way), and recent technological developments have been

based on methods and techniques that were previously not available. Some contemporary debates in AI research in general and the ethics of AI in particular will be about the methods used to achieve different kinds of functionalities in AI technologies. Such contemporary debates can, of course, be hard to keep up with.

Then again, at the same time, a case like "agentic AI" is something that also seems to pick up on much older ideas as well. Back in the 1990s, Stuart Russell and Peter Norvig even suggested that we define the whole project of creating artificial intelligence as the project of creating artificial agents that can autonomously perform tasks (Russell and Norvig 1995). And Wiener's aforementioned early statement of the AI alignment problem back in 1960 was formulated in terms of the idea of "agency" when he wrote that:

> "If we use, to achieve our purposes, a mechanical *agency* with whose operation we cannot interfere effectively . . . we had better be quite sure that the purpose put into the machine is the purpose which we really desire." (Wiener 1960, p. 1358, emphasis added)

So talk of AI as a form of agents/agency is not new, but of course recent technological developments involve new innovations and so on that help to enable what was previously primarily a theoretical idea (Kasirzadeh and Gabriel 2025). The point here, to repeat, is that there is actually a lot of continuity in theoretical discussions about the ethics of AI, so that many of the topics that are now made pressing by recent technological developments have already been part of ethical debates about AI before the technologies envisioned were possible in practice. Yet, we do of course acknowledge that the ethics of AI is a constantly developing field and what follows represents part, but not the whole of the ethics of AI. One cannot cover everything in one book, not even in a book as long as this one.

Without any further ado, let us now jump right in and start with our first debate.

As mentioned, the first topic that will be considered is the question of how to understand what AI is and, sticking with the topic of agency that was just mentioned, whether the idea of agency should be part of the definition of what AI is.

References

Altman, S. (2025). Reflections. Sam Altman Blog, 5 January. https://blog.samaltman.com/reflections (accessed 12 August 2025).

Bender, E. M., Gebru, T., McMillan-Major, A. et al. (2021). On the dangers of stochastic parrots: can language models be too big? *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. New York: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

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

Brown, P., Duffy, C., and Dubnow, S. (2025). This man says ChatGPT sparked a "spiritual awakening." His wife says it threatens their marriage. *CNN*, 2 July. https://edition.cnn.com/2025/07/02/tech/chatgpt-ai-spirituality (accessed 12 August 2025).

Chalmers, D. (2023). Could a large language model be conscious? *Boston Review*, 9 August. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/ (accessed 12 August 2025).

Coeckelbergh, M. (2020). *AI Ethics*. Cambridge, MA: MIT Press.

Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press.

Duffy, C. (2024). "There are no guardrails." This mom believes an AI chatbot is responsible for her son's suicide. *CNN*, 2 July. https://edition.cnn.com/2025/07/02/tech/chatgpt-ai-spirituality (accessed 12 August 2025).

Eadicicco, L. (2025). Anthropic says its new AI model can work almost an entire workday straight. *CNN*, 22 May. https://edition.cnn.com/2025/05/22/tech/ai-anthropic-claude-4-opus-sonnet-agent (accessed 12 August 2025).

El Atillah, I. (2023). Man ends his life after an AI chatbot '"encouraged"' him to sacrifice himself to stop climate change. *Euronews*, 31 March. https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate- (accessed 12 August 2025).

European Union. (2024). "Regulation (EU) 2024/1689" of the European Parliament and of the Council, of 13 June 2024, laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union L Series: ELI, 1–144. http://data.europa.eu/eli/reg/2024/1689/oj. (accessed 12 August 2025).

Floridi, L. (2023). *The Ethics of Artificial Intelligence – Principles, Challenges, and Opportunities*. Oxford: Oxford University Press.

Hanna, A. and Bender, E.M. (2025). *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. New York: Harper Collins.

Hashim, S. (2024). Anthropic has hired an "AI welfare" researcher. *Transformer*, 31 October. https://www.transformernews.ai/p/anthropic-ai-welfare-researcher (accessed 12 August 2025).

Heritage, S. (2025). "I felt pure, conditional love": the people who marry their AI chatbots. *The Guardian*, 12 July. https://www.theguardian.com/tv-and-radio/2025/jul/12/i-felt-pure-unconditional-love-the-people-who-marry-their-ai-chatbots (accessed 12 August 2025).

Hill, K. (2025). "She is in love with ChatGPT." *New York Times*, 15 January. https://www.nytimes.com/2025/01/15/technology/ai-chatgpt-boyfriend-companion.html (accessed 12 August 2025).

Hinton, G. (2025). Banquet speech. *The Nobel Prize*, 10 December. https://www.nobelprize.org/prizes/physics/2024/hinton/speech/ (accessed 12 August 2025).

Jefferson, G. (1949). The mind of mechanical man. *British Medical Journal* 1 (4616): 1105–1110.

Kasirzadeh, A. and Gabriel, I. (2025). Characterizing AI agents for alignment and governance. arXiv:2504.21848. https://doi.org/10.48550/arXiv.2504.21848.

Kim, J. (2025). Family shows AI video of slain victim as an impact statement – possibly a legal first. *NPR*, 12 May. https://www.npr.org/2025/05/07/g-s1-64640/ai-impact-statement-murder-victim (accessed 12 August 2025).

Minsky, M. (1968). *Semantic Information Processing*. Cambridge, MA: MIT Press.

Müller, V.C. (accepted for publication). Ethics of artificial intelligence and robotics. In: *The Stanford Encyclopedia of Philosophy (Fall 2025 Edition)* (ed. E.N. Zalta and U. Nodelman). https://plato.stanford.edu/archives/fall2025/entries/ethics-ai/.

Nyholm, S. (accepted for publication-a). Online relationships. In: *The Oxford Handbook of the Philosophy of Personal Relationships* (ed. M. Betzler and S. Stroud). Oxford: Oxford University Press.

Nyholm, S. (accepted for publication-b). *The Ethics of Artificial Intelligence: A Philosophical Introduction*. Indianapolis: Hackett.

Popper, K. (1945/2002). *The Open Society and Its Enemies*. London: Routledge.

Russell, S. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Hoboken, NJ: Prentice Hall.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences* 3 (3): 417–424.

Singleton, T., Gerken, T., and McMahon, L. (2023). How a chatbot encouraged a man who wanted to kill the Queen. *BBC*, 6 October. https://www.bbc.com/news/technology-67012224 (accessed 12 August 2025).

Suikkanen, J. (2014). *This is ethics: an introduction*. Oxford: Wiley-Blackwell.

Weber-Guskar, E. (2022). Reflecting (on) Replika: can we have a good affective relationship with a social chatbot? In: *Social Robotics and the Good Life* (ed. J. Loh and W. Loh), 103–126. Bielefeld: transcript.

Wiener, N. (1960). Some moral and technical consequences of automation. *Science* 131 (3410): 1355–1358.

van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics* 1 (3): 213–218.

Zerilli, J., Danaher, J., Maclaurin, J. et al. (2021). *A Citizen's Guide to Artificial Intelligence*. Cambridge, MA: MIT Press.