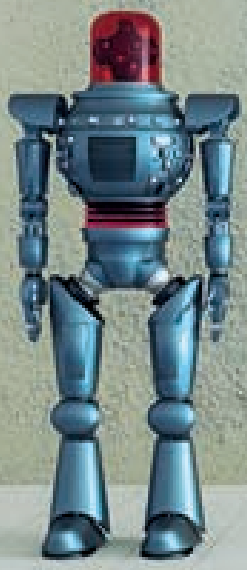
OXFORD

AI Narratives

A History of Imaginative Thinking about Intelligent Machines



Edited by
STEPHEN CAVE
KANTA DIHAL
SARAH DILLON

AI NARRATIVES

AI NARRATIVES

A History of Imaginative Thinking about Intelligent Machines

Edited by

Stephen Cave, Kanta Dihal, Sarah Dillon

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2020

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2019952162

ISBN 978–0–19–884666–6

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Contents

Part II Modern and Contemporary

Notes on Contributors

Olivia Belton is a postdoctoral research associate at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge. She recently completed a PhD on posthuman women in science fiction television at the University of East Anglia. She is currently researching public perceptions of and media representations of autonomous flight.

Stephen Cave is Executive Director of the Leverhulme Centre for the Future of Intelligence, Senior Research Associate in the Faculty of Philosophy, and Fellow of Hughes Hall, all at the University of Cambridge. After earning a PhD in philosophy from Cambridge, he joined the British Foreign Office, where he spent ten years as a policy adviser and diplomat, before returning to academia. His research interests currently focus on the nature, portrayal, and governance of AI.

Kate Devlin is Senior Lecturer in Social and Cultural Artificial Intelligence at King's College London. Her research in Human–Computer Interaction and Artificial Intelligence investigates how people interact with and react to technologies, both past and future. She is the author of *Turned On: Science, Sex and Robots* (Bloomsbury, 2018), which examines the ethical and social implications of technology and intimacy.

Kanta Dihal is a postdoctoral researcher at the Leverhulme Centre for the Future of Intelligence, University of Cambridge. She is the Principal Investigator on the Global AI Narratives project, and the Project Development Lead on Decolonizing AI. In her research, she explores how fictional and nonfictional stories shape the development and public understanding of artificial intelligence. Kanta's work intersects the fields of science communication, literature and science, and science fiction. She is currently working

on two monographs: *Stories in Superposition* (2021), based on her DPhil thesis, and *AI: A Mythology*, with Stephen Cave.

Michael Dillon is Emeritus Professor of Politics at Lancaster University. Inspired by Continental Philosophy and the intersection of modern politics, science, and religion, his work has focused on Security, on how Security has become a generative principle of formation for modern politics, and on the biopoliticization of Security in the digital age. His last book was entitled *Biopolitics of Security: A Political Analytic of Finitude* (2015). His current book project is entitled *Making Infinity Count*. Michael Dillon coedits the *Journal for Cultural Research* (https://www.tandfonline.com/rcuv20, Routledge), and is coeditor of a monograph series on *Political Theologies* for Bloomsbury (https://www.bloomsbury.com/uk/series/political-theologies/).

Sarah Dillon is a feminist scholar of contemporary literature, film and philosophy in the Faculty of English, University of Cambridge. She is author of *The Palimpsest: Literature, Criticism, Theory* (2007); *Deconstruction, Feminism, Film* (2018); and co-author (with Claire Craig) of *Storylistening: Narrative Evidence and Public Reasoning* (2021). She has edited *David Mitchell: Critical Essays* (2011); and co-edited *Maggie Gee: Critical Essays* (2015) and *Imagining Derrida* (2017), a special issue of *Derrida Today*. She is the general editor of the Gylphi Contemporary Writers: Critical Essays book series, and also broadcasts regularly on BBC Radio 3 and BBC Radio 4.

Ben Halliburton is a doctoral candidate at Saint Louis University. He is currently working on his dissertation, *'Utriusque illorum illustrium regum, pari gradu consanguineus': The Marquisate of Montferrat in the Age of Crusade, c.1135–1225*, under the supervision of Thomas Madden.

Minsoo Kang is an associate professor of history at the University of Missouri—St. Louis. He is the author of *Sublime Dreams of Living Machines: The Automaton in the European Imagination* (2011, Harvard

University Press) and *Invincible and Righteous Outlaw: The Korean Hero Hong Gildong in Literature, History, and Culture* (2018, University of Hawaii Press). He is also the translator of the Penguin Classic edition of the classic Korean novel *The Story of Hong Gildong*, and the author of the short story collection *Of Tales and Enigmas* (2006, Prime Books).

Kevin LaGrandeur is a professor of English at the New York Institute of Technology, and a Fellow of the Institute for Ethics and Emerging Technology. He specializes in literature and science, digital culture, and Artificial Intelligence and ethics. His writing has appeared in both professional venues, such as *Computers and the Humanities*, and *Science Fiction Studies*, and in the popular press, such as the *USA Today* newspaper. His books include *Artificial Slaves* (Routledge, 2013), which won a 2014 Science Fiction and Technoculture Studies Prize, and *Surviving the Machine Age* (Palgrave Macmillan, 2017).

Genevieve Liveley is a Turing Fellow and Reader in Classics at the University of Bristol, where her particular research interests lie in narratives and narrative theories (both ancient and modern). She has published widely in books, articles, and essays on narratology, on chaos theory, cyborgs, AI, and how ancient myth might help us to better anticipate the future.

Paul March-Russell is Lecturer in Comparative Literature at the University of Kent, Canterbury. He is the editor of *Foundation: The International Review of Science Fiction* and commissioning editor of the series, SF Storyworlds (Gylphi Press). His most recent book is *Modernism and Science Fiction* (Palgrave, 2015), whilst other relevant publications have appeared in *The Cambridge History of the English Short Story* (2016), *The Cambridge History of Science Fiction* (2019), *The Edinburgh Companion to the Short Story in English* (2018), and *Popular Modernism and Its Legacies* (Bloomsbury, 2018). He is currently working on contemporary British women's short fiction and animal/ecocritical theory.

Graham Matthews is an assistant professor in English at Nanyang Technological University, Singapore. His most recent book is *Will Self and Contemporary British Society* and his work on twentieth-century literature has appeared in journals and edited collections such as *Modern Fiction Studies*, *Textual Practice*, *Journal of Modern Literature*, *Critique*, *English Studies*, *Literature & Medicine*, and *The Cambridge Companion to British Postmodern Fiction.*

Anna McFarlane is a British Academy Postdoctoral Fellow at Glasgow University with a project entitled 'Products of Conception: Science Fiction and Pregnancy, 1968–2015'. She has worked on the Wellcome Trust-funded Science Fiction and the Medical Humanities project and holds a PhD from the University of St Andrews in William Gibson's science fiction novels. She is the coeditor of *Adam Roberts: Critical Essays* (Gylphi, 2016) and *The Routledge Companion to Cyberpunk Culture* (2019).

Julie Park is a scholar of seventeenth- and eighteenth-century material and visual culture working at the intersections of literary studies, information studies and textual materialism. She is Assistant Curator/Faculty Fellow at the Bobst Library of New York University, author of *The Self and It* (Stanford University Press, 2010), and co-editor of *Organic Supplements* (University of Virginia Press, 2020). Her current projects are *My Dark Room*, a study of the camera obscura as a paradigm for interiority in eighteenth-century England's built environments and *Writing's Maker*, an examination of self-inscription technologies and their materials in the long eighteenth century.

Gabriel Recchia is a research associate at the Winton Centre for Risk and Evidence Communication at the University of Cambridge, where he is currently studying how best to communicate information about risks, benefits, statistics, and scientific evidence. Previously, he was at the Centre for Research in the Arts, Social Sciences and Humanities, where he developed techniques for the analysis of large corpora of historical texts. He also

conducts research on statistical models of language and their applications in the cognitive and social sciences.

Beth Singler is the Junior Research Fellow in Artificial Intelligence at Homerton College, Cambridge. In her anthropological research she explores popular conceptions of AI and its social, ethical, and religious implications. She has been published on AI apocalypticism, AI and religion, transhumanism, and digital ethnography. As a part of her public engagement work, she has produced a series of short documentaries on AI, and the first, *Pain in the Machine*, won the 2017 AHRC Best Research Film of the Year award. Beth is also an Associate Research Fellow at the Leverhulme Centre for the Future of Intelligence.

Will Slocombe is Senior Lecturer in the Department of English at the University of Liverpool, the Director of the MA pathway in Science Fiction Studies, and one of the directors of the Olaf Stapledon Centre for Speculative Futures. He teaches modern and contemporary literature and literary theory, with a focus on science fiction. His main current research is concerned with representations of AI, and his second monograph, *Emergent Patterns: Artificial Intelligence and the Structural Imagination*, is forthcoming in 2020.

Sam Thomas completed her PhD in Classics and Ancient History at the University of Bristol, exploring the narrative ethics of Lucan's epic Civil War.

E. R. Truitt is an associate professor of Comparative Medieval History at Bryn Mawr College (Pennsylvania, USA) and the author of *Medieval Robots: Mechanism, Magic, Nature, and Art* (University of Pennsylvania Press, 2015), as well as the author of numerous scholarly articles on the history of automata and clock making, pharmacobotany, and *materia medica*, astral science, and courtly technology in the medieval period. She has written for *Aeon*, *The TLS*, and *History Today*, and she consulted on the Science Museum's exhibit "You, Robot."

Megan Ward is an assistant professor of English at Oregon State University and the author of *Seeming Human: Artificial Intelligence and Victorian Realist Character* (The Ohio State University Press, 2018). In addition, she codirects *Livingstone Online* (www.livingstoneonline.org), a digital archive of the Victorian explorer David Livingstone.

Acknowledgements

This book originates in the AI Narratives project, a joint initiative of the Leverhulme Centre for the Future of Intelligence at the University of Cambridge and the Royal Society. In 2017 and 2018, this project held four workshops to explore how AI is portrayed, what impact that has, and what we can learn from how other emerging technologies have been communicated. Each workshop convened a highly interdisciplinary group, including not only academics from a wide range of fields, but also representatives of industry, government, news media, and the arts. The participants were united by a sense of the urgent importance of engaging critically with the narratives surrounding AI at this crucial historical moment when the technology itself is advancing so rapidly. The energy and insight these events generated inspired us to prepare this collection. We would like to thank the Cambridge and Royal Society teams who prepared these workshops, and all those who participated so enthusiastically. In particular, we would like to thank Claire Craig, then Chief Science Policy Officer at the Royal Society, for her crucial role in supporting the AI Narratives project.

We would also like to thank the many people who helped in the preparation of this volume. First, the work of our research assistant, Clementine Collett, was invaluable, particularly in the final weeks before the submission deadline. We are immensely grateful to those colleagues who reviewed our (the editors') contributions to this collection: Olivia Belton, Rachel Adams, Will Slocombe, and Michael Shapiro. We would also like to thank the team at Oxford University Press, who have so smoothly guided this project to publication, in particular Ania Wronski, Francesca McMahon, and Sonke Adlung. Finally, we acknowledge the generous support of the Leverhulme Trust, through their grant to the Leverhulme Centre for the Future of Intelligence, and the support of the University of Cambridge Faculty of English's Research Support Fund.

Introduction

Imagining AI

Stephen Cave, Kanta Dihal, and Sarah Dillon

> 'I mean, those parables or whatever they are,
> maybe they mean a lot to you but, uh . . .'
> (Sladek 1980, p.52)

0.1 AI Narratives

In 1985, David Pringle included John Sladek's *Roderick*—the story of a robot who wanders across a near-future America—in his list of the one hundred best science fiction novels. Pringle declares the novel, and its sequel *Roderick at Random* (1983), 'a treatise on the whole theme of mechanical men, homunculi, automatons and machine intelligence—the ultimate robot novel' (1985). The novel opens at a brilliantly complex moment of crisis built around a situation no doubt familiar to many artificial intelligence (AI) researchers—the question of funding. A small lab in a little-heard-of university in a near future where everything has been automated—from the grading of university papers to police detection—has been receiving financial support from NASA for an AI research project. Only it turns out that the whole funding setup has been a scam designed to line the pockets of a NASA employee who commits suicide on being exposed. Nevertheless, the academics have made breathtakingly exciting progress with the research and have created Roderick, 'a learning system'

(Sladek 1980, p.24) described by one character as 'this artificial intelligence' (p.23). 'He's alive,' insists one of the researchers to his colleague, 'Roderick's alive. I know he's nothing, not even a body, just content-addressable memory. I could erase him in a minute—but he's alive. He's as real as I am [...]. He's realer. I'm just one of his thoughts' (p.48). The research team is now in the unenviable position of having to appeal to the University's Emergency Finance Committee in order to continue with the research, but not all the committee members are persuaded by the value of their work.

Rogers, a sociologist on the committee, visits the robotics lab to see what it is all about, but is disappointed by what he finds: 'a lot of computers and screens and things, I could see those anywhere, and what are they supposed to mean to a layman? I expected—I don't know—' (Sladek 1980, p.24). He arrives with preconceptions about what he would find, preconceptions easily deduced by the AI researcher, Fong:

> 'You wanted a steel man with eyes lighting up? "Yes Master?", that kind of robot? Listen, Roderick's not like that. He's not, he doesn't even have a body, not yet, he's just, he's a learning system [...] A learning system isn't a thing, maybe we shouldn't even call him a robot, he's more of a, he's like a *mind*. I guess you could call him an artificial mind.' (Sladek 1980, pp.24–25)

Rogers scoffs at the esoteric nature of Fong's claim, and at the absence of the embodied AI he is expecting to see: 'am I supposed to tell the committee I came to see the machine and all you could show me was the ghost?' (Sladek 1980, p.25). But it turns out Rogers is less interested in the artificial mind than he is in the mind of the researcher, quizzing Fong about 'this Frankenstein goal' (p.25) and whether he has ever considered 'the social impact of your work' (p.26).

Roderick is laden with references to the AI narratives that have preceded it, from Kurt Vonnegut's dystopia of automation, *Player Piano* (1952), to the myth of Francis Bacon's brazen head, to

Albertus Magnus' automaton, to 'the prophet Jeremiah and his son, making the first *golem*' (Sladek 1980, p.52). Dr Jane Hannah, an anthropologist on the Emergency Finance Committee, evidences a detailed knowledge of AI narratives across the globe. In fact, she turns to these traditions in order to try to understand the desires behind the aspiration to create intelligent machines, and to decide whether she should approve funding the continued development of one at her university:

> '... maybe the Blackfeet boy, Kut-o-yis, cooked to life in a cooking pot, but isn't that the point? Aren't they always fodder for our desires? Take Pumiyathon for instance, going to bed with his ivory creation [...] take Hephaestus then, those golden girls he made who could talk, help him forge, who knows what else ... Or Daedalus, not just the statues that guarded the labyrinth, but the dolls he made for the daughters of Cocalus, you see? Love, work, conversation, guard duty, baby, plaything, of course they used them to replace people, isn't that the point? [...] And in Boeotia, the little Daedala, the procession where they carried an oaken bride to the river, much like the *argeioi* in Rome, the puppets the Vestal Virgins threw into the Tiber to purge the demons; disease, probably, just as the Ewe made clay figures to draw off the spirit of the smallpox, so did the Baganda, they buried the figures under roads and the first [...] person who passed by picked up the sickness. In Borneo they drew sickness into wooden images, so did the Dyaks [...] Of course the Chinese mostly made toys, a jade automaton in the Fourth Century but much earlier even the first Han Emperor had a little mechanical orchestra [...] but the Japanese, Prince Kaya was it? Yes, made a wooden figure that held a big bowl, it helped the people water their rice paddies during the drought. Certainly more practical than the Chinese, or even the Pythagoreans, with their steam-driven wooden pigeon, hardly counts even if they did mean it to carry souls up to – but no, we have to make do with the rest, and of course the golem stories, and how clay men fashioned by the Archangel – [...] There were the Teraphim of course, but no one knows their function. But the real question is, what do we want this robot *for*? Is it to be a bronze Talos, grinning as he clasps people in his red-hot

> metal embrace? Or an ivory Galatea with limbs so cunningly jointed – ' (Sladek 1980, pp.60–61)

Sladek's novel understands the tradition it lies within, that is, a transhistorical, transcultural imaginative history of intelligent machines. These imaginings occur in a diverse range of narrative forms, in myths, legends, apocryphal stories, rumours, fiction, and nonfiction (particularly of the more speculative kind). They have existed centuries prior to the origin of the modern scientific field, which might most simply be located in 1956 at the Dartmouth Summer Research Project on Artificial Intelligence, the term 'AI' having been coined the year prior (McCarthy et al 1955). The term is now used to refer to a heterogeneous network of technologies—including machine learning, natural language processing, expert systems, deep learning, computer vision, and robotics—which have in common the automation of functions of the human brain.[1] However, imaginings of intelligent machines have employed a range of other terms, including 'automaton' (antiquity), 'android' (1728), 'robot' (1921), and 'cyborg' (1960).[2] The chapters in this book engage with AI in its broadest sense, one that encompasses all of the aforementioned terms: that is, any machine that is imagined as intelligent.

The exploration of AI narratives by Dr Hannah in *Roderick* covers imaginings from the literature, mythology, and folklore of Native America, Ancient Greece, Classical Rome, Uganda, Ghana, Borneo, China, Japan, Judaism, and Christianity. The chapters in this book focus on narratives that form part of the Anglophone Western tradition. These include works written in English and works in other languages that have had a strong influence on this narrative tradition. The chapters therefore cover a historical period beginning with the automata of the *Iliad*—the oldest narrative of intelligent machines in this tradition, written around 800 BCE—to the present. The book presents this history of imaginative thinking about intelligent machines in two parts. The chapters in Part I cover the long history of imaginings of

AI from antiquity to modernity. Each chapter in this part focuses on a specific historical period: antiquity, the Middle Ages, the Renaissance, the eighteenth century, the nineteenth century, and the modernist period. Together, they explore the prehistory of key concerns of contemporary AI discourse, including: the nature of mind; the imbrication of AI and power; the duality of our fascination with and yet ambivalence about AI; the rights and remuneration of workers (both human and artificial); the relation between artificial voice and intelligence; creativity; and technophobia. Part II takes up the historical account in the modern period, Karel Čapek's 1921 play *R.U.R.* (from which derives the term 'robot') serving as the hinge between the two parts. The chapters in Part II focus on the twentieth and twenty-first centuries, in which a greater density of narratives emerges alongside rapid developments in AI technology. The chapters in this part are organised thematically, with each chapter driven primarily by a focus on imaginative explorations of specific effects and consequences of AI technologies. These include: the dehumanizing effects of humanizing machines; the consequences of automation and mechanization for society; the cultural assumptions embedded in the anthropomorphization of machines; the importance of understanding AI as a distributed phenomenon; the human drivers behind the desire for technologically enabled immortality; the interaction of AI with the sovereign-governance game that defines modern rule; the relationship between imaginative representations of female robots and AIs, and the perception of real-world sex robot technology; and the fear of losing control of AI technologies. This book covers many of the touchstone narratives that might be most familiar to readers, including Isaac Asimov's robot stories and *The Terminator*, but it also aims to draw readers' attention to less well-known texts, such as *Roderick*, that make an important contribution to the rich imaginative history of intelligent machines.

After her exploration of AI narratives, Dr Hannah comes to a conclusion regarding how she will vote at the Emergency Finance Committee: 'As you see,' she concludes, 'I've been turning the

problem over, consulting the old stories. . . . And I've decided to vote against this robot' (Sladek 1980, pp.60–61). For Dr Hannah, AI narratives offer complex explorations of the social, ethical, political, and philosophical consequences of AI, explorations that inform her decision-making about whether or not to fund contemporary scientific research. Many of the chapters in this book support this position, exploring how AI narratives have addressed, and offer sophisticated thinking about, some of the legitimate concerns that AI technologies now raise. But AI narratives are not universally viewed in this way. A counter-narrative is found in *Roderick* in the form of Dr Hannah's colleague, Dr Helen Boag. Her position cuts through the ancient myths and the expectations they create, ones she perceives to be often misaligned with the technology itself: 'really isn't the computer more or less an overgrown adding machine? A tool, in other words, useful of course but only in the hands of human beings. I feel the role of the computer in our age has been somewhat exaggerated, don't you?' (Sladek 1980, p.73). Other chapters in this book engage with the challenges posed by, primarily, dominant AI narratives. These engage with a wider landscape in which prevalent AI narratives are mistrusted or criticised for example for their extremism—utopian or dystopian—or for their misrepresentation of current technology, for instance in their tendency to focus on anthropomorphic representations. In the next section of this introduction, we want to survey that wider landscape of contemporary views about prevalent AI narratives, their functions and effects, as well as the impacts they have, in order to map one of the terrains into which we hope this book will intervene.

0.2 The Impact of Narratives

Sheila Jasanoff's concept of 'sociotechnical imaginaries' provides a dominant paradigm for understanding the relationship between technology and the social order. Jasanoff defines these as the

'collectively held, institutionally stabilized, and publicly performed visions of desirable futures, animated by shared understandings of forms of social life and social order attainable through, and supportive of, advances in science and technology' (2015, p.4). Absent from this theorisation, however, is an explicit account of the important role narratives play, both fictional and nonfictional, as fundamental animators of sociotechnical imaginaries. This neglect is at odds with Jasanoff's use of science fiction narratives, for instance, to introduce her theory, and her acknowledgement that science fiction narratives already provide an established site of investigation of sociotechnical imaginaries, long prior in fact to the invention of the term: science fiction already 'situate[s] technologies within [...] integrated material, moral, and social landscapes [...] in such abundance' (p.3). The work of this book contributes to establishing the importance of narratives as constituent parts of any sociotechnical imaginary. Attention to narratives also highlights relationships between society, technology, and the imaginary that are not included in Jasanoff's definition, for instance, visions that are not collectively held, ones that destabilise institutions, and subaltern narratives. Jasanoff acknowledges that sociotechnical imaginaries encompass both 'positive and negative imaginings' (p.3), but only in service of the dominant vision. Foregrounding narratives therefore plays a role in challenging the 'aspirational and normative' (p.5) dimension of Jasanoff's concept, inviting consideration of a much wider range of visions.

Narratives of intelligent machines matter because they form the backdrop against which AI systems are being developed, and against which these developments are interpreted and assessed. Those who are engaged with AI either as researchers or regulators are therefore rightfully concerned, for instance, about the fact that the dominant contemporary imaginings of AI, primarily those of Hollywood cinema and popular news coverage, are often out of kilter with the present state of the technology. The UK House of Lords Select Committee on Artificial Intelligence opens

the second chapter of their 2018 report 'AI in the UK: ready, willing and able?' with a sharp critique of prevalent AI narratives:

> The representation of artificial intelligence in popular culture is lightyears away from the often more complex and mundane reality. Based on representations in popular culture and the media, the non-specialist would be forgiven for picturing AI as a humanoid robot (with or without murderous intentions), or at the very least a highly intelligent, disembodied voice able to assist seamlessly with a range of tasks. (p.22)

The Select Committee report continues by observing that 'many AI researchers were concerned that the public were being presented with overly negative or outlandish depictions of AI, and that this could trigger a public backlash which could make their work more difficult' (p.24).

This book contributes to an emerging body of work that goes beyond hype and horror by exploring the more complex ways in which narratives of AI could have significant impact. For instance, the narratives with which AI researchers themselves engage can influence their 'career choice, research focus, community formation, social and ethical thinking, and science communication' (Dillon & Schaffer-Goddard, forthcoming). Within these categories, Dillon and Schaffer-Goddard identify that further investigation is needed into the way in which AI narratives might influence who goes into the field. Scholars such as Alison Adam have been considering for over two decades how masculinity is inscribed into the way AI is conceived (1998) and how this might interplay with a culture in the computing world that is hostile to women. Future research might consider, for instance, whether the consistent portrayal of fictional AI developers as men, from Hephaestus to *Metropolis*'s (1927) Rotwang to Robert Ford of the *Westworld* TV series (2016–present), makes women feel that this role is not for them. This question of who is in the room, or who is in the lab, impacts which systems are developed, how, and for whom.

The dominance of anthropomorphic portrayals of AI also exacerbates the tense relationship between the technology and

issues of equality and diversity. Anthropomorphic machines, from fictional androids like the 'hosts' of *Westworld* to the human voices of virtual personal assistants, are mirrors to our societies, perpetuating existing biases and exclusions. They reflect and so reinforce prevalent cultural narratives of different groups' allotted role and worth, delineating further who counts as fully human (Rhee 2018).[3] Extending work on AI narratives to social justice issues around race and ethnicity is crucial. Further research here might consider, for instance, whether stock images of AI as Caucasian male humanoid robots distort views of who AI is for, and whose jobs will be impacted by this technology, obscuring the potential effect on disadvantaged communities.

In order to understand what kinds of technologies are being developed, those outside the professional AI field rely on narratives that mediate between the technology world and the public sphere. Narratives can therefore strongly influence public acceptance and uptake of AI systems, and a significant amount of science popularisation has this goal explicitly in mind (Gregory 2003). The Select Committee report notes that the role of AI narratives as an intermediary between research and the public was a concern raised by AI researchers, who 'told [the Committee] that the public have an unduly negative view of AI and its implications, which in their view had largely been created by Hollywood depictions and sensationalist, inaccurate media reporting' (Select Committee on Artificial Intelligence 2018, p.44). This view is supported by Dillon and Schaffer-Goddard's findings, who argue that 'such stories have a strong influence on the researchers' science communication activity—researchers often need to argue against the pictures they paint, but can also use this as an incentive and a springboard to paint more positive, or at least realistic, pictures of AI-influenced futures'. They found that AI researchers expressed 'a strong desire for more sophisticated stories about AI, which would be of benefit to the research community and public discourse, as well as to literary and cinematic quality and production' (Dillon & Schaffer-Goddard, forthcoming). Many of the chapters

in this book identify such stories, encouraging attention to them, in contrast to the narratives that currently dominate.

Prevalent narratives are of course not just those of the popular media and the press. Large corporations such as Microsoft and Google invest significant resources in developing ethics principles and other narratives aimed at fostering public acceptance of AI. Whether this technology is adopted has implications far beyond these companies' bottom lines. Use of AI in healthcare for instance, will impact the advancement of the medical field and individual well-being and mortality rates. At the same time, these technologies pose significant concerns regarding privacy, social justice, workers' rights, democracy, and more. In this context, AI narratives have played, and will continue to play, a crucial role in determining the future of AI implementation.

By influencing the perceptions of policymakers, and by steering public concerns, narratives also affect the regulation of AI systems. For example, there is ongoing debate about the ways narratives influence public policy on highly advanced or superintelligent AI (Johnson & Verdicchio 2017). On the one hand, there are those who argue that current real-world risks are being obscured by Terminator-style stories. The Select Committee report notes that prevalent AI narratives 'were concentrating attention on threats which are still remote, such as the possibility of "superintelligent" artificial general intelligence, while distracting attention away from more immediate risks and problems' (Select Committee on Artificial Intelligence 2018, p.23). In contrast, there is the potential for those who oppose restraints on the development of superintelligent AI to propagate narratives sceptical of its capacities, so that policymakers see no need for regulation (Baum 2018).

Many authors and scholars consciously use narratives to explore the possibilities for a future with intelligent machines and disrupt existing tropes. Imaginative thinking about AI can probe both dystopian and utopian scenarios, showing flaws in overly unidirectional thinking, and anticipating consequences before they

affect the lives of millions of people. Asimov, for instance, developed his robot stories in response to a plethora of science fiction works in the 1920s and 30s that presented, in his view, unhelpful depictions of robots as either extremely menacing or extremely pathetic characters (1995 [1982]). Currently, initiatives around the world are producing a growing body of work—especially science fiction—that is commissioned with the specific mandate to explore the impact of particular technologies, or to imagine how current technologies can be developed to create a utopian future (Amos & Page 2014; Coldicutt & Brown 2018; Finn & Cramer 2014). At the same time, the field of futures studies uses narratives to explore the consequences of specific decisions in policymaking and scientific development (Avin 2019; B. D. Johnson 2011). Their scenarios are intended to be realistic forecasts directing the focus of AI ethics research.

The way AI is portrayed is therefore a social, ethical, and political issue. Through shaping the technical field, the acceptance of the resulting technology, and its regulation, and through encoding normative sociopolitical assumptions, these portrayals have far-reaching implications. It is therefore essential that prevalent narratives be critically examined, and that they be contested by the privileging of more sophisticated and complex narratives of AI, both fictional and nonfictional. These more complex stories can be and are being newly invented, but this book draws attention to the extensive history of imaginative thinking about intelligent machines upon which they might build, and by which contemporary thinking about AI should be informed. This book looks to past imaginings—of the future and of alternate realities—in order to inform present thinking about AI.

0.3 Chapter Guide

Philosopher Daniel Dennett offers one illustration of the impact of AI narratives on the scientific field. In 1997, Dennett wrote an eight-page fan letter to Richard Powers after reading *Galatea 2.2*

(Powers 1995). Explaining his actions later, Dennett observes of Powers's novel:

> His representation of AI is wonderful. It is remarkable how this very interested bystander has managed to cantilever his understanding of the field out over the abyss of confusion and even throw some pioneering light on topics I thought I understood before. (2008, pp.151–52)

For Dennett, Powers's novel serves to illuminate the field: 'What particularly excited me,' he continues, 'was how Powers had managed to find brilliant ways of conveying hard-to-comprehend details of the field, details people in the field were themselves having trouble getting clear about' (Dennett 2008, p.152). Dennett understands Powers' novel as a contribution to scientific research and knowledge. 'The novel,' he says, 'is an excellent genre for pushing the scientific imagination into new places' (p.160). This is because it challenges 'personal styles of thinking, and a style is (roughly) a kit of *partially disabling* thinking-habits' (p.160). Dennett is keen to caveat this acknowledgement of the power of novels, however, declaring that 'you have to be a powerful thinker to pull off the trick. That's why most science fiction doesn't repay the attention of scientists' (p.161). He even goes so far as to suggest that we need a new name 'for the rare novels like *Galatea 2.2* that manage to make a contribution to the scientific imagination' (p.161). No new name is needed, however, and *Galatea 2.2* is not so rare a contribution as Dennett seems to think. The chapters in this book explore a wide range of AI narratives and the many roles they play not only in extending the scientific imagination, but the ethical, political, and social imagination as well.

Part I opens with a chapter on the very earliest references to intelligent machines. Through close analysis of Homer's *Iliad* and *Odyssey*, Genevieve Liveley and Sam Thomas trace the various gradations of weak to strong machine 'intelligence' that these ancient poems describe, and explore the ancient mind models that they assume. They conclude that the Homeric mind

model—which sees both humans and machines possessing programmable thoughts (*phrenes*) and minds (*noos*)—helps to explain why Homer is part of our own cultural AI programming. They propose that Homer's ancient tales of intelligent machines have established a 'programme' for us to follow in the retelling and rescripting of our present and future AI narratives. Part I then moves from antiquity to the long medieval period. First, E. R. Truitt explores the imagination of objects we might recognize as AI: from artificial servants to elaborate optical devices, from augmented human perception to sentient machines. She argues that throughout this era, AI appears in imaginative contexts to gain, consolidate, and exercise power. It does so in ways remarkably recognizable from a twenty-first-century standpoint, from maintaining class and gender hierarchies, to aiding military or political dominance, to gaining knowledge of the future that could be used to a person's advantage. In the next chapter, Minsoo Kang and Ben Halliburton focus on this question of foreknowledge by examining the major appearances of the story of the speaking head of philosopher Albertus Magnus (c. 1193–1280) in medieval and early modern writings. The head is a wondrous object able to converse and even reason, confirming it as a kind of medieval AI, animated for the purpose of divination. Kang and Halliburton demonstrate that, in a way that is reflective of contemporary anxiety and concerns about AI, the different versions of the story move between questioning whether Albertus was dealing in illicit knowledge in making the object, secularizing it as a purely mechanical device, or demonizing it as a work involving diabolical beings. The changing narratives around the head exhibit both a fascination with, as well as an anxious ambivalence towards, the object, revealing that our current attitudes towards AI originate from long-standing and primordial feelings about artificial simulacra of the human.

In the next chapter, Kevin LaGrandeur engages with a different talking brass head, this one built by a natural philosopher in Robert Greene's Elizabethan comedy *Friar Bacon and Friar Bungay*

(1594). Engaging with Greene's play as well as Renaissance stories of the golem of Prague and of Paracelsus's homunculus, LaGrandeur demonstrates the fears and hopes embedded in Renaissance culture's reactions to human invention. In particular, his chapter exposes the entanglement of AI narratives with discourses of slavery, showing how intelligent objects of that period are almost uniformly proxies for indentured servants. The tales with which LaGrandeur engages signal ambivalence about our innate technological abilities—an ambivalence that long predates today's concerns. The promise represented by these artificial servants, of vastly increased power over natural human limits, are countervailed by fears about being overwhelmed by our own ingenuity. Julie Park shifts attention from the artificial head to the artificial speech that emanates from it, investigating what the eighteenth-century history of artificial voices tell us about the relationships between machines, voice, the human, and fiction. Given the rise in our daily lives of voice-operated 'intelligent assistants', this investigation is especially pertinent. By examining the eighteenth-century case of a speaking doll and the cultural values and desires that its representation in a 1784 pamphlet entitled *The Speaking Figure, and the Automaton Chess-Player, Exposed and Detected* reveals, Park's chapter provides a historical framework for probing how the experiences and possibilities of artificial voice shed light on our deep investments in the notion of voice as the ultimate sign of being 'real' as humans.

In designing his foundational test of AI, Alan Turing refers to Ada Lovelace's Victorian pronouncement that a machine cannot be intelligent because it only does what it is programmed to do. This idea continues to shape the field of computational creativity as the 'Lovelace objection'. In her chapter, Megan Ward argues that this term is a misnomer and sets out to show that Ada Lovelace actually proposes a much more nuanced understanding of human-machine collaboration. By situating Lovelace's work within broader Victorian debates about originality in literary realism, especially in relation to Charles Dickens and Anthony

Trollope, Ward demonstrates how Lovelace's ideas participate in a broader Victorian debate that redefines originality to include technologically enhanced mimesis. Ward proposes that understanding computational creativity not in opposition to the Lovelace objection, then, but as the development from Victorian originality to contemporary creativity, may open a way forwards to a concept of creativity inclusive of mechanicity. Paul March-Russell's chapter concludes Part II's historical sweep from antiquity to the present with a focus on the technophobia of modernist literature towards the question of machine intelligence. Ranging across texts by Edmund Husserl, Albert Robida, Emile Zola, Ambrose Bierce, Samuel Butler, H. G. Wells, E. M. Forster, and Karel Čapek, March-Russell demonstrates that, whilst modernism may have been enthusiastic towards other forms of technological innovation, the possibility of a machine that could think—and, above all, talk—like a human stirred age-old responses about the boundaries between human and nonhuman life-forms. March-Russell concludes that, despite Čapek's popularization of the term 'robot', it is only with the rise of science fiction that literary authors begin to supersede the technophobia of their modernist predecessors.

Čapek's play provides the hinge from Part I, covering antiquity to modernity, to Part II, which focuses on modern and contemporary AI narratives, predominantly, although not exclusively, science fictional. In her chapter, Kanta Dihal engages with Karel Čapek's *R.U.R.* (1921), alongside Ridley Scott's *Blade Runner* (1982) and Jo Walton's Thessaly trilogy (2014–2016), in order to contextualize the imagining of the robot uprising in fiction against the long history of slave revolts. Doing so, she shows that in these three works, the revolt is depicted as a justified assertion of personhood by the intelligent machines. These stories show how intelligent machines are just as likely to be denied personhood as historically oppressed groups have been: those in power are unwilling to grant personhood if this were to threaten their personal comfort. In the next chapter, Will Slocombe also draws

together texts from the early twentieth century to the present day. Slocombe engages with four different imagined 'Machines' in E. M. Forster's 'The Machine Stops' (1909); Paul W. Fairman *I, The Machine* (1968); Asimov's 'The Evitable Conflict' (1950); and the popular television series *Person of Interest* (2011–2016). He examines the ways in which representations of AI during the twentieth century—particularly in nonandroid form (what might be termed a 'distributed system')—dovetail with pre-existing perceptions of society operating as a 'machine', and become imbricated in broader social discourses about loss of autonomy and individuality.

Graham Matthews' chapter focuses on the mid-twentieth century, a crucial period for the technological development of AI, the moment at which, it might be said, AI moved from the realm of myth to become a real possibility. Matthews contends that midcentury AI narratives must be situated in relation to concomitant technological developments in automation and cybernetics. These elicited widespread concern among government institutions, businesses, and the public about the projected transition from industry to services, the threat of mass unemployment, technologically driven sociopolitical change, and the evolving relationship between humans, religion, and machines. The rhetoric of the Fourth Industrial Revolution closely echoes these midcentury debates. Matthews analyses the varied representation of AI in midcentury novels such as Michael Frayn's *The Tin Men* (1965); Len Deighton's *Billion Dollar Brain* (1966); and Arthur C. Clarke's *2001: A Space Odyssey* (1968). He argues that these novels problematize AI narrative tropes by resisting anthropomorphic tendencies and implausible utopian and dystopian scenarios; instead, they address the societal ramifications—both positive and negative—for humans faced with technological breakthroughs in AI.

Beth Singler's chapter moves the historical focus into the second half of the twentieth century and introduces a series of new methodological approaches in the book's later chapters. Singler employs historical and cognitive anthropological approaches to examine the cultural influences on our stories

about AI, focusing in particular on anthropomorphization, and attending to the cultural assumptions about the human child, and the AI 'child', present in our AI narratives. She works her exploration through engagement with a range of texts, including the novel *When HARLIE Was One* (1972); the films *D.A.R.Y.L.* (1985), *AI: Artificial Intelligence* (2001), *Star Trek: Insurrection* (1998), and *Tron: Legacy*, (2010); the television series *Star Trek: Next Generation* (1987 to 1994); and a speculative nonfiction account of the future relationship of humanity and AI, Hans Moravec's *Mind Children* (1988). Anna McFarlane approaches AI narratives through the lens of genre, focusing on cyberpunk science fiction and the possibilities opened, both in fictional narratives and wider cultural narratives, by the genre's sustained interrogation of AI as a phenomenon that is dispersed throughout networks. She argues that the narrative innovations and tropes of early cyberpunk writers such as William Gibson and Samuel R. Delany, and the continued mutation of these ideas in 'post-cyberpunk', such as in the work of Cory Doctorow, respond to exponentially changing technology, and shape contemporary understandings of AI's cutting edge, for instance algorithmic decision-making. Stephen Cave explores representations of immortality in science fiction texts primarily from the cyberpunk tradition—works by William Gibson, Greg Egan, Pat Cadigan, Robert Sawyer, Rudy Rucker, and Cory Doctorow. These authors offer subtle, frequently sceptical portrayals of the psychological, philosophical and technological challenges of using technology to free oneself from the constraints of the body. Cave shows that these works offer a particularly important site of critique and response to techno-utopian narratives by influential contemporary technologists—in particular Hans Moravec and Ray Kurzweil.

In their chapter, Sarah Dillon and Michael Dillon bring political theory into dialogue with literary criticism in order to explore the interaction between AI and the ancient conflict between sovereignty and governance, in which sovereignty issues the warrant to rule and governance operationalizes it. They focus on three novels in which games, governance, and AI weave themselves

through the text's fabric: Iain M. Banks's *The Player of Games* (1988) and *Excession* (1996), and Ann Leckie's *Ancillary Justice* (2013). These novels play out the sovereign-governance game with both artificial and human actors. In doing so, Dillon and Dillon argue that the novels question what might be politically novel about AI, but in the end, reveal that whilst AI impacts the pieces on the board, reducing some and advancing others, it does not materially change the logic of the game. They conclude that these texts therefore raise questions but do not provide answers with regard to what might be required for AI technologies to change the algorithms of modern rule. Moving from the politics of sovereignty and governance to the politics of gender, Kate Devlin and Olivia Belton's chapter explores the relationship between fictional representations of female robots and AIs and the perception of real-world sex robot technology. They offer a critical analysis of science fiction media representations of gendered AI, focusing on Ava from *Ex Machina* (2014), Samantha from *Her* (2013), and Joi from *Blade Runner 2049* (2017). They show how these AIs' gender identities are often reinforced in stereotypical ways, and how both embodied and disembodied AIs remain highly sexualized. Their chapter demonstrates how discourses around real-life sex robots are deeply informed by prevalent fictional narratives, and advocates for a more gender-equal approach to the creation of both fictional and factual robots, in order to combat sexist stereotypes.

The final chapter of the collection introduces a digital humanities methodology to approaching AI narratives, that is, one that combines the tools of traditional literary analysis with computational techniques for identifying key themes and trends in very large quantities of text. Gabriel Recchia presents a computationally assisted analysis of the English-language portion of the Open Subtitles Corpus, a dataset of over 100,000 film subtitles ranging from the era of silent film to the present. By applying techniques used to understand large corpora within the digital humanities, Recchia presents a qualitative and quantitative overview of several salient themes and trends in the way AI is portrayed and discussed

in twentieth- and twenty-first-century film. Recchia's analysis confirms many of the dominant themes and concerns discussed in the previous chapters of the book, whilst also opening the way for new ways of approaching and analysing AI narratives in the future.

0.4 Conclusion

Through our personal engagement with contemporary debates on the impact of AI, the three of us have witnessed many Roderick moments in the last few years. Misplaced expectations and mutual incomprehension between stakeholders have been as much a part of the present moment's AI revolution as extravagant utopian and dystopian visions. In a recent survey on perceptions of this technology, when asked how the respondent would explain AI to a friend, several responses were simply expressions of anxiety, such as 'creepy' or 'scary robots' (Cave, Coughlan, & Dihal 2019). Despite these concerns, it is likely that AI technologies will be highly consequential for the shape of society in the near and long term. If their effects are to be positive rather than negative, it will be essential to reconcile the multiple discourses of different publics, policymakers, and technologists, and lay bare the assumptions and preconceptions on which they rest. We hope this book, in beginning to unpick the fascinating and complex history of AI narratives, will contribute to that goal.

Notes

1. This definition of 'AI' is informed by Marcus Tomalin's introductory talk at the workshop 'The Future of Artificial Intelligence: Language, Gender, Technology', 17 May 2019, University of Cambridge.
2. These various terms and their related imaginaries have been explored in a rich body of scholarly work with which this collection is in dialogue, and to which it aims to contribute. Several of the contributors to this collection have written key works in the field, including Minsoo Kang (2011), Kevin LaGrandeur (2013), E. R. Truitt (2015), and Megan Ward (2018).

3. Clementine Collett and Sarah Dillon's report 'AI and Gender: Four Proposals for Future Research' (2019) outlines the challenges AI technologies, including humanoid robotics and virtual personal assistants, present to gender equality; identifies current research and initiatives; and proposes four areas for future research.

Reference List

Adam, A. (1998) *Artificial knowing: gender and the thinking machine*. London, Routledge.

Amos, M. & R. Page (eds.) (2014) *Beta Life: short stories from an a-life future*. Manchester, Comma Press.

Asimov, I. (1995 [1982]) Introduction. In: *The complete robot*. London, HarperCollins. pp.9–12.

Avin, S. (2019) Exploring artificial intelligence futures. *Journal of AI Humanities*. https://doi.org/10.17863/cam.35812.

Baum, S. (2018) Superintelligence skepticism as a political tool. *Information*. 9(9), 209. https://doi.org/10.3390/info9090209.

Cave, S., K. Coughlan, & K. Dihal (2019) 'Scary robots': examining public responses to AI. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.17863/CAM.35741.

Coldicutt, R. & S. Brown (eds.) (2018) *Women invent the future*. Available from: https://doteveryone.org.uk/project/women-invent-the-future/ [Accessed 18 September 2019].

Collett, C. & S. Dillon (2019) *AI and gender: four proposals for future research*. Cambridge, Leverhulme Centre for the Future of Intelligence.

Dennett, D. (2008) Astride the two cultures: a letter to Richard Powers, updated. In: S. J. Burn & P. Dempsey (eds.) *Intersections: essays on Richard Powers*. Champaign, IL, Dalkey Archive Press. pp.151–61.

Dillon, S. & J. Schaffer-Goddard (forthcoming). What AI researchers read: the influence of stories on artificial intelligence research.

Finn, E. & K. Cramer (eds.) (2014) *Hieroglyph: stories and visions for a better future*. New York, William Morrow.

Gregory, J. (2003) Understanding 'science and the public': research and regulation. *Journal of Commercial Biotechnology*. 10(2), 131–39.

Jasanoff, S. (2015) Future imperfect: science, technology, and the imaginations of modernity. In: S. Jasanoff & S.-H. Kim (eds.), *Dreamscapes of modernity: sociotechnical imaginaries and the fabrication of power*. Chicago: University of Chicago Press. pp.1–33.

Johnson, B. D. (2011) Science fiction prototyping: designing the future with science fiction. *Synthesis Lectures on Computer Science*. 3(1), 1–190. https://doi.org/10.2200/S00336ED1V01Y201102CSL003.

Johnson, D. G. & M. Verdicchio (2017). Reframing AI discourse. *Minds and Machines*. 27(4), 575–90. https://doi.org/10.1007/s11023-017-9417-6.

Kang, M. (2011) *Sublime dreams of living machines: the automaton in the European imagination*. Cambridge, MA, Harvard University Press.

LaGrandeur, K. (2013) *Androids and intelligent networks in early modern literature and culture: artificial slaves*. New York, Routledge.

McCarthy, J., M. L. Minsky, N. Rochester, & C. E. Shannon (1955) *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. 31 August. Available from: http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf [Accessed 18 September 2019].

Powers, R. (1995) *Galatea 2.2*. New York, Picador.

Pringle, D. (1985) *Science fiction: the 100 best novels* (Kindle). London, Science Fiction Gateway.

Rhee, J. (2018) *The robotic imaginary: the human and the price of dehumanized labor*. Minneapolis, University of Minnesota Press.

Select Committee on Artificial Intelligence (2018) *AI in the UK: ready, willing, and able?* (No. HL 100 2017–19). Available from: House of Lords website: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf [Accessed 18 September 2019].

Sladek, J. (1968) *The reproductive system*. London, Granada.

Sladek, J. (1980) *Roderick, or, the education of a young machine*. New York, Carroll & Graff.

Sladek, J. (1983) *Roderick at random*. London, Granada.

Truitt, E. R. (2015) *Medieval robots: mechanism, magic, nature, and art*. Philadelphia, University of Pennsylvania Press.

Ward, M. (2018) *Seeming human: Victorian realist character and artificial intelligence*. Columbus, Ohio State University Press.

PART I

ANTIQUITY TO MODERNITY

1

Homer's Intelligent Machines

AI in Antiquity

Genevieve Liveley and Sam Thomas

1.1 Introduction

Intelligent machines have been a staple of narrative fiction for the past 3,000 years and, in the classical Greek and Roman myth kitty, we find a significant corpus of stories featuring devices which exhibit varying degrees of artificial intelligence (see Mayor 2018; Bur 2016; Liveley 2005; Liveley 2019; Rogers & Stephens 2012; Rogers & Stephens 2015). Representing the earliest phase of this ancient literary tradition, the poet Homer describes relatively simple mechanical devices such as self-pumping bellows (*Iliad* 18.468) and self-opening gates (*Iliad* 5.748–52 and 8.392–6) that appear able to anticipate the desires of their users and perform basic repetitive tasks spontaneously and with a moderate degree of autonomy. These 'almost intelligent widgets' (Pfleeger 2015, p.8) exhibit what Gasser and Almeida characterize as 'weak (or narrow)' rather than 'strong (or general) AI' (2017, p.59).[1] However, Homer also describes slightly more complex contraptions that exhibit correspondingly stronger and more developed levels of (quasi) intelligent automation: multipurpose tripods—serving as tables, altars, and stands—that are represented as *automatos* or 'self-acting' (*Iliad* 18.373–19.379) as they move back and forth between the homes of the gods.[2] Yet more sophisticated automata are deemed

to possess something in addition to this power simply to anticipate their users' needs and to move 'of themselves' or 'without visible cause'—as Homer typically describes the operations of these devices. In Homeric epic, we also find a higher order of 'intelligent' machines whose narrative representation suggests that something more significant and cognitively complex than rudimentary automation is being imagined. Homer apparently refers to a pair of silver and gold watchdogs which guard the palace of Alcinoos (whose name actually means 'Strong mind') not with sharp teeth and claws but with their own supposedly 'intelligent minds' (*Odyssey* 7.91–7.94). The ships that eventually take Odysseus home to Ithaca not only move as 'fast as a thought' (*Odyssey* 7.36) but 'navigate by thought' (*Odyssey* 8.556). And the slaves made of gold who serve as personal assistants to the god Hephaestus (*Iliad* 18.418–18.422) exhibit a still higher order of intelligence: they not only possess human form but have the power of movement, of speech, and of thought too. These machines have voices, physical strength, and—uniquely among Homer's wondrous machines—intelligent *minds*.

Through close literary analysis of these Homeric devices, tracing the various gradations of weak to strong machine 'intelligence' that the epic poems describe and the mind models that these gradations assume, this chapter considers what these ancient narratives might tell us about the ancient history of AI. Beginning with a re-examination of Homer's weak AI, his simple automata and autonomous vehicles, in order to provide context and so help us better to appreciate the more sophisticated models of artificial mind and machine cognition attributed to Homer's stronger, embodied AI, this chapter asks: What kinds of priorities and paradigms do we find in AI stories from Homeric epic and (how) do these still resonate in contemporary discourse on AI? In particular, what (if any) distinctions does Homer draw between artificial and human minds and intelligences? And what (if any) is the legacy of Homer's intelligent machines and the ancient narrative history of AI?

1.2 Homer's Automata

Although not *sensu stricto* intelligent machines, Homer's eighth-century-BCE descriptions of self-opening gates (*Iliad* 5.748–5.752 and 8.392–8.396), self-pumping bellows (*Iliad* 18.468–18.473), and self-propelling tripods (*Iliad* 18.373–18.379) provide an important background measurement against which to take our reading of Homer's more sophisticated AIs.

The Homeric gods do not keep slaves for manual work (Garlan 1988, p.32) but could hardly be expected to open and close their own gates, so Homer grants them a pair of automatic gates to their heavenly citadel which 'of themselves groan on their hinges' (*automatai de pulai mukon*) as they spontaneously open for the goddess Juno and her chariot at the very same moment as she touches her horses with a whip (*Iliad* 5.748–5.749; the same formula is repeated at 8.393). Similarly, the metalworking god Hephaestus does not keep slaves to work his furnace, but he has an automated machine which controls the variable intensity of the heat supply to his crucibles, wherein he can thereby simultaneously smelt bronze, tin, silver, and gold (*Iliad* 18.468–18.473). Although these bellows are not explicitly characterized by Homer as *automatai*, the fact that there is a total of twenty in operation here, servicing four separate smelting processes, indicates that the two-handed, club-footed god is not pumping them all himself. In such a context, and at such a semi-industrial scale, these devices are evidently working at some level—like the gates of heaven—autonomously. What is more, we are told that the bellows work not in response to Hephaestus' manual pumping but at his command—he *orders* them to work (*Iliad* 18.469)—and that they vary their outputs according to his wishes/instructions, to suit what Hephaestus desires (*Iliad* 18.473). We are not told *how* the bellows might 'know' what Hephaestus wants or needs, or how his commands are communicated, received, and processed. However, the key verb used to describe the object of Hephaestus' commands (*ergazesthai*) is revealing here: it is a term typically used

by Homer and his contemporaries to refer to the manual labour of slaves. Hera apparently opens the automated gates of heaven with the crack of her horse whip, and Hephaestus engages with this automated machine as if it were his slave.

The bellows, like the gates of heaven, are represented as tools that replace human slaves and, as such, are assumed to function both mechanically and cognitively on a basic level akin to that of their human counterparts. In this context—reflecting an ancient culture in which slavery was widespread—Homer and his audiences would readily understand that the bellows are able to 'know' what their user desires them to do in the same way that a slave is able to 'know' what its master wants it to do. The user master gives an order and the machine slave obeys. The human (or, in this case, the divine) user master demonstrates his capacity for higher-order cognitive function in his powers of judgement, of technical expertise, of decision-making, and the like; the machine slave demonstrates its lesser cognitive capacity in doing what it is told. User master and machine slave think and work separately, on different cognitive planes, yet synergetically, towards the same goals.

This same synergetic—and hierarchical relationship—is also suggested in Homer's description of the thing that user master and machine slave are together employed in producing here. For Homer's Hephaestus uses one of these basic-model machine slaves to aid him in the task of manufacturing automata of even greater technical ingenuity and (quasi) intelligence. For Hephaestus uses his self-pumping and self-regulating bellows to make a set of self-moving tripods (*Iliad* 18.372–18.381):

> He moved to and fro about his bellows in eager haste; for he was manufacturing tripods (*tripodas*), twenty in all, to stand around the wall of his well-built hall. He had set golden wheels (*kukla*) on to the base of each one so that of themselves (*automatoi*) they could enter the assembly of the gods for him/at his bidding (*hoi*) and return again to his house, a wonder to see (*thauma idesthai*). They were almost fully finished, but the clever/cunning (*daidalea*)

> handles/ears (*ouata*) were not yet fixed upon them. He was making these now, and was cutting the rivets (*kopte de desmous*), working away with intelligent understanding (*iduiesi prapidessi*), . . .

The connection between the bellows and the artefacts they produce is reinforced here by the reference to their same generous number: there are twenty bellows (*Iliad* 18.468–18.473) powering the furnace that Hephaestus is using to manufacture these twenty tripods (*Iliad* 18.374). And, just as the bellows represent both ingenious product and process, these tripods are explicitly represented by Homer as the products of Hephaestus' mechanical prowess—objects which demonstrate and even share in his technological ingenuity, here characterized as his 'clever' or 'cunning skill' (*daidalea*).[3]

Homer gives us a relatively detailed picture of how the tripods are fashioned: golden wheels (*kukla*) are fixed to the base of each tripod, and their elaborate handles or 'ears' (*ouata*) are attached with metal rivets. Yet Homer tells us relatively little about their operation. Again, like the automated bellows which work at Hephaestus' command (*Iliad* 18.469) and 'in whatever way' required (*Iliad* 18.473), the tripods are also supposed to move 'at [Hephaestus'] bidding (*hoi*)' (*Iliad* 18.376). However, Homer does not spell out for us what mind model might enable these devices (with their 'clever' ears) to know or anticipate what Hephaestus wants or needs. Crucially, nor does he tell us what the tripods are supposed to do. The practical functionality of the automatic doors and automated bellows was clear. What is less clear from Homer's narrative is why Hephaestus' self-moving tripods, these *automatoi*, are enhanced not only with golden wheels (in and of itself a marker of relative technological sophistication) but with the power to move both of their own accord and at their master's bidding. Why are these further enhancements desirable in this context?

Some of this uncertainty arises from the ambiguity of the artefact. Tripods are regularly featured in Homeric epic and appear to have had a range of purposes, reflecting their status as high-value