INTRODUCTION

A Philosophical Thinking Machine

Over the course of its brief seventy-year history the field of artificial intelligence (AI) has known a succession of "golden ages" during which advances are rapidly made, and "ice ages" when progress has disappointingly slowed. Most commentators would agree that we are currently in the midst of one of these AI golden ages. Since the success of Deepmind's *AlphaGo* program against Go champion Lee Seedol in 2016 neural networks and deep learning have rarely been out of the news. The following claims about the limits of these newly fashionable forms of artificial intelligence were recently posted on the internet:

> Artificial intelligence programs like deep learning neural networks may be able to beat humans at playing Go or chess, or doing arithmetic, or writing Navy Seal copypasta, but they will never be able to truly think for themselves, to have consciousness, to feel any of the richness and complexity of the world that we mere humans can feel. Mere, unenlightened humans might be impressed by the abilities of simple deep learning programs, but when looked at in a more holistic manner, it all adds up to… well, nothing. They still don't exhibit any trace of consciousness. All of the available data support the notion that humans feel and experience the world differently than computers do. While a computer can beat a human master at chess or Go or some other game of structured rules, it will never be able to truly think outside of those rules, it will never be able to come up with its own new strategies on the fly, it will never be able to feel, to react, the way a human can. Artificial intelligence programs lack consciousness and self-awareness. They will never be able to have a sense of humor. They will never be able to appreciate art, or beauty, or love. They will never feel lonely. They will never have empathy for other people, for animals, for the environment. They will never enjoy music or fall in love, or cry at the drop of a hat. Merely by existing, mere, unenlightened humans are intellectually superior to computers, no matter how good our computers get at winning games like Go or Jeopardy. We don't live by the rules of those games. Our minds are much, much bigger than that.

The possibility or otherwise of computer consciousness has been much-debated and it remains a controversial topic—so there is little that's remarkable about the claims being made in this passage. What is more remarkable is *who* wrote it: the passage was composed in its entirety *by* a computer, OpenAI's GPT-3. It so happens that GTP-3 is itself a neural network-type system, one that possesses an internal model of the English language

comprising some 175 billion parameters, powered by deep learning algorithms and trained by exposure to the entirety of the internet and libraries of books.[1]

Anyone conducting a broader survey of GPT-3's outputs—in addition to philosophy it is able to produce include poetry, conversations, songs, jokes, legal prose and restaurant menu—will quickly discover that the program is far from infallible, and the mistakes that it makes suggest that it lacks anything resembling a full understanding of what it is writing about. The machine's linguistic skills are enviable, but it falls short of possessing the level of wide-ranging general intelligence that we possess. But as David Chalmers has suggested:

> Nevertheless, GPT-3 is instantly one of the most interesting and important AI systems ever produced. This is not just because of its impressive conversational and writing abilities. It was certainly disconcerting to have GPT-3 produce a plausible-looking interview with me. GPT-3 seems to be closer to passing the Turing test than any other system to date (although "closer" does not mean "close") …
>
> More remarkably, GPT-3 is showing hints of general intelligence. Previous AI systems have performed well in specialized domains such as game-playing, but cross-domain general intelligence has seemed far off. GPT-3 shows impressive abilities across many domains. It can learn to perform tasks on the fly from a few examples, when nothing was explicitly programmed in. It can play chess and Go, albeit not especially well. Significantly, it can write its own computer programs given a few informal instructions. It can even design machine learning models. Thankfully they are not as powerful as GPT-3 itself (the singularity is not here yet).**[2]**

With advances such as these being made it is not surprising that in recent years increasing numbers of people have begun to take seriously the idea that artificial intelligence that rivals or even surpasses that of human beings is a genuine possibility, and are pondering their implications.

From Animal Souls to Machine minds and the Turing Test

They may be much in the recent news, but the issue of whether or not an artificial construct can possess a life or mind of its own is by no means a new one. Thinkers in earlier centuries were well aware that this issue has the potential to have an enormous impact: on how we should think of ourselves and what our place in the universe really is. If we could build a machine that has the same sort of mental capacities as a human being, then we humans can't be as special as many of us would like to think.

---

[1] GPT-3 stands for "generative pre-trained transformer version three", and it has been exposed to approximately 45 billion times the number of words an average human being sees in their entire life. For further details about how the cited text was generated see **[1, 13.7]**

To anyone with a passing acquaintance with the history of Western philosophy René Descartes (1596 – 1650) is a familiar figure—indeed, Descartes is often referred to as "the father of modern philosophy".[2] He is famous (at least among philosophy undergraduates) for wondering in his *Meditations* whether he could possibly be certain—absolutely certain—that he was not dreaming or being deceived by an evil demon. Irrespective of his paternal relationship with modern philosophy, Descartes did have a real daughter, Francine, who sadly died of scarlet fever in 1640 when only five years old. After Descartes' own death a strange rumour started to spread to the effect that Descartes had constructed a fully life-like automaton that closely resembled in appearance his daughter Francine—the doll was said to accompany Descartes everywhere on his travels. On one of these trips a ship's captain is alleged to have accidentally opened the case where the automaton was stored, and horrified by what he found cast it into the sea.

The full story of how this rumour originated is a fascinating and complex one, but there can be little that doubt it was often passed on with a view to discrediting Descartes and his followers, some of whom were associated with then-scandalous forms of materialism.[3] For present purposes it provides a useful illustration of just how controversial some of Descartes' views were. In the 17th and 18th centuries the issue of the extent to which human beings are nothing more than purely physical machines was giving rise to increasingly heated debates, and Descartes views were central to these debates.

One of Descartes' more infamous doctrines was his stance on the sort of minds possessed by non-human animals. Referring to the latter as "bête machines" he denied that they have conscious mentality. If you step on a puppy's tail it may well squeal and bark, but you can reassure yourself that it is not feeling any pain. Explaining his views to Henry More he wrote: "The greatest of all the prejudices we have retained from infancy is that of believing that brutes think." **[4, 544]** Few contemporary philosophers find Descartes' stance on animal minds plausible, and even in his day it had comparatively few takers.[4] However, the reasons Descartes put forward for adopting this stance are of considerable contemporary relevance.

---

[2] He is also a familiar character in contemporary philosophy of mind texts for defending a form of dualism, holding that our minds reside not in our brains, but in immaterial soul-substances. While the typical undergraduate textbook portrait of Descartes is not entirely misleading, it is also guilty of concealing the true scope of intellectual endeavours. While his philosophy was certainly important to him, Descartes devoted more time and effort on mathematics, physics and biology, and his writings in the latter fields were influential in the 17th and 18th centuries. If "Cartesian Dualism" features in any dictionary of philosophy, "Cartesian coordinates" (also invented by Descartes) will feature in any dictionary of mathematics – and most of us will have encountered them at school.

[3] For the full story of Descartes' robotic daughter, in all its fascinating complexity, see **[3]**.

[4] Catherine Descartes, the philosopher's niece, observing that a female warbler bird returning to the same window year after year remarked to a friend "with all due respect to my uncle, she has judgement." See **[5, 75]** for further details.

In the 17th century, Descartes' view that we humans possess a soul was wholly unremarkable—at the time, everyone (or nearly everyone) would have agreed. In contrast, Descartes' claim that animals are nothing more than machines was far from commonplace—At the time it was quite revolutionary. The dominant world-view at the time was Aristotelian, and for Aristotelians the world was chock-full of souls of one kind or another. Plants were thought to possess a *nutritive soul* responsible for their basic life-functions, and which allowed them to feed, grow and reproduce. Animals possessed a nutritive soul, but in addition they possessed a *sensitive* soul, which allows them to perceive their surroundings and move their bodies. Human beings possessed nutritive and sensitive souls, but they also possess a *rational soul*, responsible for their distinctive intellectual capacities.

Like other forward-looking thinkers during the early phases of the scientific revolution Descartes was eager to abolish any trace of (to his eyes near magical) Aristotelian souls from the material world. Consequently he held that all physical things—even highly complex ones such as plants and animals—are constituted entirely of material parts that are governed by simple mechanical laws. These material parts are invisibly small, and the laws governing them are akin to the laws of motion governing observable things such as thrown balls passing through the air, pendulums and the inner mechanical workings of clocks. It is these mechanical laws—rather than anything resembling Aristotelian animistic souls—that are responsible for all aspects of plant and animal life. While it is uncontroversially the case that living things such as roses, oak trees, frogs, birds and dogs *appear* very different from mechanical objects such as clocks or musical boxes, for Descartes these appearances are deceptive: in fact, they are all fundamentally of the same nature, living things are nothing more than complex material mechanisms.

We now know that Descartes was correct—certainly his mechanical view of living things is one nearly all contemporary biologists would accept. However, this victory did not occur overnight. The doctrine that living things are special and fundamentally different from the non-living still seemed plausible to many scientists in the 19th century, and it was only with advances in molecular biology in the early decades of the 20th century that it was finally put to bed. Given all this, is scarcely surprising that so many in the 17th century found Descartes' mechanical view of life so shocking and absurd.[5]

What of human beings? Descartes was one of the leading biologists of his day and being well-versed in the theory and practice of dissection. Given that he was fully aware that similar structures can be found within the brains and bodies of human and animals it was not surprising to find that argued that our own bodies are also machines. Descartes held that all the basic operations of a human body could be fully explicated in mechanical terms, without any need for the nutritive and sensory souls posited by the Aristotelians. However, there was

[5] Even writing a full century after Descartes, when La Mettrie published his *L'homme machine* in 1747 readers found it so outrageous that La Mettrie had to flee the usually very tolerant Netherlands.

one aspect of human life that Descartes could not conceive a mere machine being capable of replicating: our reason or intelligence. In his *Discourse on the Method* (1637) Descartes wrote:

> … if any such machines had the organs and outward shape of a monkey or of some other animal that doesn't have reason, we couldn't tell that they didn't possess entirely the same nature as these animals; whereas if any such machines bore a resemblance to *our* bodies and imitated as many of *our* actions as was practically possible, we would still have two very sure signs that they were nevertheless not real men. **(1)** The first is that they could never use words or other constructed signs, as we do to declare our thoughts to others. We can easily conceive of a machine so constructed that it utters words, and even utters words that correspond to bodily actions that will cause a change in its organs (touch it in one spot and it asks 'What do you mean?', touch it in another and it cries out 'That hurts!', and so on); but *not* that such a machine should produce different sequences of words so as to give an appropriately meaningful answer to *whatever* is said in its presence—which is something that the dullest of men *can* do. **(2)** Secondly, even though such machines might do some things as well as we do them, or perhaps even better, they would be bound to fail in others; and that would show us that they weren't acting through understanding but only from the disposition of their organs. For whereas reason is a universal instrument that can be used in all kinds of situations, these organs need some particular disposition for each particular action; hence it is practically impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way our reason makes us act. **[6, p.22]**

So far as Descartes was concerned, no purely mechanical system could possibly possess the ability to converse on any and all topics in the way we effortlessly do at a moment's notice. Nor could such a machine find solutions to an indefinitely wide range of problems in the way that we manage to do—human intelligence is a "universal instrument". It was this stance on the ultimate limitations of physical machinery that led Descartes to conclude that the rational parts of our minds could not be physical.

By virtue of being non-physical, an immaterial soul is free from the limitations governing physical machines. If nothing physical could possess our intellectual capacities, these capacities must reside in something non-physical, and an immaterial soul is the obvious candidate. Hence while Descartes found that he could dispense with two of the Aristotelian souls, he felt obliged to retain a version of the rational soul.[6] Since the behavioural repertoire

[6] Some of Descartes' contemporaries were more radical and were prepared to reject his dualistic conception of human beings entirely. In his *Leviathan* (1651) Thomas Hobbes maintained that there is no human capacity that is incapable of being explained in mechanical material terms. Margaret Cavendish also found dualism problematic and argued for an all-inclusive materialism: "I would ask those, that say the Brain has neither sense, reason, nor self-motion, and therefore no Perception; but that

of non-human animals is far less complex—they can't converse and can only solve a narrowly circumscribed range of problems—Descartes saw no obstacle to regarding *them* as purely physical machines, devoid of the immaterial soul that we possess.

In 1950 Alan Turing published "Computing Machinery and Intelligence" in the philosophy journal *Mind* **[8]**. Here Turing proposed his famous and much-discussed test for machine intelligence. If a computer could be programmed so as to replicate the conversational skills of an average human being by providing appropriate and meaningful responses to whatever is put to it then it would be legitimate to regard the computer as possessing genuine intelligence. Turing's test is clearly anticipated in the passage of *Discourse* cited above. Descartes may have thought it unlikely that a wholly physical machine could replicate the intelligent behaviour of a human being, but he also seemed willing to accept that if this were to occur it would be legitimate to regard the machine as being genuinely intelligent and a rational agent.

From the standpoint of our technologically sophisticated 21st century we should certainly be wary of being overly critical of Descartes and his views as to the feats ordinary physical machinery might be capable of achieving. After all, the most advanced technologies in his the day were spring-powered clocks and the water-powered automata that could be found in gardens of the richer members of the nobility.[7] If he had lived to see billions of transistors being crammed onto small computer chips would he have adopted a different stance? Would he have been even more impressed when he learned that Turing had proved that these machines have the very special power to compute everything that is mathematically computable? We can only speculate, but it is by no means impossible.

Questions, Issues, Problems

In his 1950 paper Turing predicted that we would not have long to wait before a computer passed his test: "I believe that at the end of the century the use of words and general educated opinion will have altered so much that the one will be able to speak of machines thinking without expecting to be contradicted." In this at least he was mistaken: by the turn of the millennium no computer had managed to pass Turing's test, and in this respect at least Descartes pessimism with regard to the potential for machine intelligence has thus far been vindicated. However, as we have seen, thanks to recent developments the prospects for genuine machine intelligence are considerably brighter than they have been for some time,

---

all proceeds from an Immaterial Principle, and an Incorporeal Sprit, distinct from the body, which moveth and actuates corporeal matter; I would fain ask them, I say, where their Immaterial Ideas reside, in what part or place of the Body?" **[7]**

[7] It should be noted that some of the automata in the Early Modern period were highly sophisticated pieces of machinery, and could seem strikingly life-like. Jacques de Vaucanson's "digesting duck', for example, had some four hundred moving parts in its wings alone. For more on relevant automata and Descartes see **[5]**.

and it may well not be very long before we have to deal with AIs that are at least as intelligent as a typical human.

In this connection there are a number of issues that have already received a good deal of attention, and which are likely to receive more in the decades to come.

One important issue concerns the relationship between a capacity for intelligent behaviour and consciousness: does genuine intelligence require consciousness? Would a machine capable of intelligent behaviour also have to be capable of having experiences of pleasure and pain, or colour and sound? Would it be capable of engaging in conscious thinking? Quite possibly, but very different views on this issue have been defended, but thus far nothing resembling a consensus has emerged. Some philosophers hold that genuine intelligence involves the capacity to consciously *understand* what one is doing, a capacity which obviously requires consciousness. But the majority of computer scientists would follow Turing's lead and reject this claim, and with some plausibility: if a computer could pass the Turing Test without being conscious, it would be odd to deny that it had a considerable degree of *some* kind of genuine intelligence. A further complicating factor here is that neuroscientists, psychologists and evolutionary biologists have found it difficult to specify with any clarity quite what explanatory role human consciousness plays in human behaviour.

A distinct but related issue concerns the very possibility of computer possessing *any* form of consciousness—the issue that was vexing GPT-3 in the passage we encountered earlier. This remains one of the most controversial questions in the philosophy of mind, and opinions remains sharply divided. For some philosophers computer consciousness is eminently possible, others rule it out as quite absurd. A complicating factor to bear in mind when considering this question is that "computers" can come in very different guises. The Turing-type that most of us are acquainted with—those found in our phones and laptops—are algorithmic devices: their program consists of a set of instructions which they carry out in a step-wise fashion. Evidently, computers are this kind are distinctly unlike biological brains, which in the human case consist of a hundred billion or so neurons, each connected to hundreds or thousands of other neurons, all working in parallel. But since the earliest days of artificial intelligence computer scientists have been designing computers that work very differently from Turing-type machines, computers which much more similar to biological brains. The "neural nets" currently associated with the revolution in machine learning fall into this category. If it should turn out that Turing-type machines are in fact the wrong kind of thing to be conscious, the same may well not be true for differently structured non-biological machines.[8]

[8] For more on these issues see the Glossary entries for "consciousness" , "consciousness: the hard problem", "Consciousness and Science Fiction" and "Cartesian Dualism".

The philosophical relevance of AI is not confined to philosophy of mind, it also gives rise to interesting ethical and political questions. If robots possessing human-level intelligence appear on the scene, how should we treat them? Should they be granted the same rights and respect as a human being? What sorts of personal relationships between AIs and humans are appropriate? Under what circumstances should you take an AI as a friend or lover? The AIs in Asimov's robot stories are programmed to "obey the orders given it by human beings, except where this would lead to a human being coming to harm". In effect if not name, Asimov's robots are slaves. Would it be morally right to create beings of this kind?[9]

A different range of pressing issues combine social, political and economic considerations. The machine intelligences available at present do not possess human-level intelligence, but they are sufficiently intelligent to do the sorts of jobs that millions of humans currently do, and as they improve they'll be able to do more. According to one recent estimate **[10]**, we can expect 35% of the workforce to be replaced by AIs over the next twenty years.[10] Predictions are of course risky, but the job widely believed to be most at risk include factory workers, lawyers, accountants and taxi drivers—and by the time GPT-5 arrives philosophers, poets and novelists might be at risk too. Working out ways of responding to these developments which maximize the potential benefits while minimizing unwanted disruption is likely will be among the greatest social and political challenges facing us over the next few decades.

Another issue that has already provoked considerable debate concern the dangers associated with the increased possibilities for mass surveillance that advances in AI are making possible. By combining data harvested from social media and internet use, location tracking via mobile phone, pervasive video surveillance cameras and facial recognition, computers capable of speedily handling vast amounts of data very quickly make it possible for interested parties to know vastly more about ordinary citizens than has hitherto been possible, and plan accordingly. Totalitarian regimes have been quick to exploit these technological possibilities, but in democratic nations too these technologies have already led to new methods for influencing the outcomes of elections—methods that unscrupulous parties have been quick to exploit, and which regulators are struggling to deal with effectively.[11]

On an economic level, the data global social media companies possess about their users has proven to be a highly saleable asset, and highly attractive to advertisers willing to

[9] See **[9]** for a useful selection of current thinking on human-robot relationships, autonomous weapons and vehicles, and a number of other issues.
[10] The website https://willrobotstakemyjob.com gives a 94% chance of accountants and auditors being replaced by AI and robots.
[11] See **[11, part III]** for Yuval Harari's perturbing reflections on the consequences of big data algorithms knowing us better than we know ourselves)

pay for it—a combination of factors which had led to social media companies accruing vast amounts of wealth in a comparatively short period of time. As both surveillance technologies and the abilities of AIs to interpret enormous quantities of data advance in the years to come, finding ways of dealing with the consequences will be a major concern.[12]

These technologies also open up new political possibilities. In China, the way the mass surveillance system is being linked to their "Social Credit" system has attracted a good deal of attention. The latter allocates penalty points to citizens who are behave in ways the state doesn't like—failing to show up for restaurant reservations, traffic violations, cheating on public transport—as well as reward points for doing things the state approves, such as donating blood or performing community services. The potential for a "Big Brother"-style micro-control of entire populations is as obvious as it is perturbing.[13] However, there is also the potential for more positive developments.

As artificial intelligence becomes more powerful it may well become possible to organize societies in ways that are simply *more intelligent* than anything presently possible. Advocates of free market capitalism are fond of claiming that if a market economic system has its downside, it is still the most efficient way possible for organizing an economy and generating wealth. No central state planner would ever be capable of monitoring the billions of economic transactions that take place on a daily basis and manage them more intelligently than the blind hand of the market. Or so a familiar line of argument runs. But even if this is the case at the moment, will it still be the case when powerful AI's that are able to exploit the resources of big data become available? Firms such as Facebook and Amazon are already monitoring billions of transactions on a daily basis, and managing them in highly effective ways. Is an AI-powered version of communism something we should dread, or look forward to? Is there any alternative to coping with the higher levels of unemployment AIs are going to produce?[14]

In the eyes of many the most important issue in this connection is working out how to protect ourselves against future machines that equal or surpass humans in intelligence. Humans are smart, but not *that* smart. It would be great to have someone a good deal smarter than us, to help solve pressing problems such as climate change, curing cancer, and reconciling quantum mechanics with general relativity—all problems which continue to defeat the most brilliant human minds. Hence there is a powerful impetus not to stop at creating AIs with human-level intelligence, but to aim for AI's that are *superintelligent*, AIs that possess vastly more intelligence than any human. But if there are lots of advantages in having a superintelligent machine at ones disposal, there are also lots of potential dangers. A

[12] An theme thoroughly explored by Shoshana Zuboff, see **[12].**
[13] For a useful overview see https://theconversation.com/chinas-social-credit-system-puts-its-people-under-pressure-to-be-model-citizens-89963
[14] See **[13]** and **[14]** for surveys of these issues.

maliciously intentioned superintelligence could decide to wipe out the human race entirely – by engineering (say) a virus with a lethality rate of 100%.

Before creating a superintelligence it would obviously be a good idea to ensure that it won't decide to do anything along these lines. But precisely what steps should we take? Given that a superintelligent machine might well be vastly smarter than us, can we be confident that *any* measure we are capable of adopting is guaranteed to succeed?[15]

From Science Fact to Science Fiction

When it comes to addressing issues such as these the science fiction genre contains an enormous quantity of resources that it would be decidedly unintelligent to ignore. Artificial intelligences—in all manner of guises and forms—has been a prominent theme in science fiction since its very beginnings. Inevitably, when devising these scenarios science fiction writers have been considering potential responses to many of the issues we have just outlined, often with considerable foresight and intelligence. In many domains the gap between science fiction and science fact is rapidly closing, and science fiction writers have long been exploring the relevant territory in interesting and thought-provoking ways—and in some cases actually helping shape it. Science fiction doesn't just have the potential to influence current thinking on AI and robotics, in many areas it has *already* exerted a very considerable influence. When writing just now about the threat that future AIs might pose to humanity, it was very difficult to avoid mentioning the *Terminator* movies or 2001's HAL. When talking of the possibility of falling in love with a machine *Blade Runner* and *Her* come quickly to mind, as does the new TV version of *Westworld.* Given the extent to which science fiction has already permeated our broader contemporary consciousness, subjecting this influence to proper scrutiny is clearly something which should be done.

Hence this book. The essays which follow explore the way minds and artificial mentality have been treated in science fiction over the past century or so, with a view to drawing out and reflecting on their implications for the issues such as those just outlined. Given the vast amount of brilliant and thought-provoking science fiction that has been produced during this period that is relevant to these topics, in what follows we have barely scratched the surface. Even so, we are confident that the essays demonstrate that the exercise is very worthwhile.

The first section, 'Qualities of Mind', explores the ways in which identity and personhood relate to AI, and how these characteristics might relate to ethics and morality. Bohn"s "Is Ava a person?" examines the extent to which Ava, from *Ex Machina* (dir. Alex Garland, 2014) might be said to be a "person" (rather than a mere "object"), and how the

[15] For more on this threat, how seriously we should take it, and what we might do to minimize the dangers, see **[15]** and **[16]**.

notion of consciousness is relevant to this issue, along with a range of other concepts, including intentionality, free will, instrumental rationality, and moral responsibility. In so doing, Bohn demonstrates that "human" and "person" are not necessarily interchangeable as categories, and highlights the extent to which personhood is defined from without (from how an entity is perceived) as much as within (whether you have the capability to see yourself as a person). Hauskeller addresses a similar question in his "What Is It Like to Be a Bot?" , albeit through authors such as Philip K. Dick, Isaac Asimov, and Ian McEwan. He considers whether many of the fears about AI are because of own our (likely justified) concerns that a system's "alternative" cognition might lead it disagree with human concerns and conclusions, and explores how empathy and emotions feed into this debate. The final chapter of the section, Slocombe & Dennis's "Governor Modules and Moral Judges", relates Martha Wells's *Murderbot Diaries* to existing research within computer science on creating ethical frameworks within AI systems, and asks how such a system might function, and how it might impact on an AI's "autonomy". Here, it is the imposition of an ethical framework on an entity and how it might govern its actions. All three chapters inflect questions about the relationship between an entity's identity in different ways, and begin to consider "relational" aspects of advanced AI and the human sphere.

These relational aspects are further examined in the second section, "Meetings of Minds". The focus here is on issues which arise when humans and AIs enter into intimate relationships, whether emotional or physical or both. We are reminded here of John McCarthy—the computer scientist who coined the term "Artificial Intelligence" as a systemic approach to computer cognition—and his short story "The Robot and the Baby" (2004), where a furious societal debate occurs about whether it is possible for a robot to love a baby. Although (it turns out) McCarthy's robot is *not* capable of loving a baby, questions about how humans might love AIs, and how AIs might love humans, recur across science fiction. Kind's essay "Love in the Time of AI" examiners *Her* (dir. Spike Jonze, 2013) alongside an episode of *Black Mirror* in order to consider how romantic love (as opposed to other kinds of love) forges a connection between human and AI, and explores the parameters of that connection. In comparison, Cave and Dihal's "AI will always love you" ranges across a wider range of science fiction, including works by Brian Aldiss and Greg Egan, and serials such as *Westworld*, *Humans* and *Real Humans*, to offer examples of three "successful" different types of love (friendship, familial and romantic), and the potential problems that emerge from human/AI relationships. Broadening out from love, Roy-Faderman's "Anne Leckie's Ancilliaries: Artificial Intelligence and Embodiment" uses texts such as the lesser-known *The Clockwork Man* by E.V. Odile (1923) and William Gibson's seminal Neuromancer (1984), alongside Leckie's recent *Ancillary* trilogy. This chapter offers a suite of ideas about the potential "emotional lives" of AI, and draws together many connection with other possible beings.

In the third section, “Changing Minds” the emphasis shifts from the personal and interpersonal aspects of AI into broader territory. Each of the four chapters here focuses on civilizational and species-level concerns and developments. Clark’s “Selfless Civilizations”, for example, is a meditation upon the fact that non-conscious machine intelligences spread widely throughout the universe, with human conceptions of “consciousness” being merely a drop in the cosmic ocean, gesturing towards authors such as Peter Watts and Charles Stross, as well as Isaac Asimov, Ray Bradbury, and C. J. Cherryh. Ćirković’s “Mindhunter: Transcending Geocentrism andPpsychocentrism in The Invincible and Peace on Earth” echoes this engagement with non-conscious intelligences, but through the filter of two of Stanislaw Lem’s novels. In much of his science fiction Lem was concerned to open us up to the possibility—or even probability—that minds elsewhere in the universe might be very different from anything human beings possess or can easily conceive. Ćirković suggests that unless and until this lesson of Lem’s is taken on board, the Copernican revolution will not be complete. Silcox’s “Historicism, Science Fiction, and the Singularity” introduces us to Karl Popper’s critique of historicism, and the considerations which led Popper to maintain that it was impossible to make reliable predictive claims about the future. As Silcox notes, recent claims concerning an allegedly imminent Singularity would be seriously undermined if Popper is right. He then moves on to explore science fiction settings created by Bruce Sterling and Iain M. Banks, and demonstrates the valuable insights these fictions provides into the intelligibility (or otherwise) of the truly advanced AIs that we might one day encounter. The final chapter, Szollosy’s “Shifting the goalposts”, brings us back to earth, but is a fitting conclusion to the collection as, through discussions of works such Asimov’s *Bicentennial Man* and *Humans*, it exposes the ways in which AI has prompted a continuous reappraisal of the “human” and how the goalposts on what we understand AI to be have shifted and continue to shift. He suggests that the necessary next step—to reconsider the frames through which we evaluate ideas such as “ethics” or “the human”—is one that will enable us to ask better questions, rather than rehearse the same old politics of exceptionalism that dominate debates about AI.

An important idea that emerges from the summaries above is that we have been deliberately broad about the very definition of the central concern of this book: what is an “artificial intelligence” anyway? Some of the contributions discuss robots, some androids/gynoids (*gendered* robots), some describe what might be termed “software agents” or programs, and still others explore cyborgs and AI/human hybrids. Moreover, some of these beings are overtly conscious, some are not conscious as we would understand the term, and some we just don’t know (and that is of course the point). As a result, some of the insights of individual discussions remain case-specific whereas others have a more general valency. But the point is that, as we are “minding the future” and being mindful of it, all of these illustrate the fact that the signifying phrase, “artificial intelligence”, is itself contested, and that different definitions of the very words will lead to markedly divergent interpretations

of what AI can and might do. Science fiction is replete with examples of all of the above, and that very proliferation can be productive in considering what an AI is, and how we might interpret it, and furthermore help us to think through how we might relate to it and how it might relate to us (with the "we" and "us" in that sentence being similarly ambiguously defined; if you're an AI reading this, who do you think "we" are…?)

What unites these varied contributions is the fact that science fiction enables varied ways of thinking about artificial intelligence and the impacts it might have. Science fiction, perhaps taking a familiar metaphor too far, operates as a kind of "simulation" of possible futures. Some of the scenarios are probable, some of them are vastly improbable. But no matter their plausibility, they can nonetheless spark ideas and approaches about the technologies that comprise AI, our attitudes towards those technologies, and the kinds of impact AI might have on us, on personal, social, global and cosmological levels. By virtue of not being confined to the present time, or constrained by current levels of technology, science fiction has the capacity to speculate about possible futures on a grander scale than other disciplines. As a consequence, science fiction has much to offer anyone interested in the large-scale picture of how conscious intelligence and the broader cosmos are related—presently and in the near and distant futures. Nowhere else is so much sheer imaginative power devoted to exploring what minds—both natural and artificial—have the potential to become. When it comes to exploring the vast space of possible minds, imagination is by far the most valuable tool we possess, and science fiction writers possess more than most.

REFERENCES

(1) GPT-3 Creative Fiction 13.7 "Why Deep Learning Will Never Truly *X*" https://www.gwern.net/GPT-3

(2) Chalmers, D.: GPT-3 and General Intelligence. *Daily Nous*, July 30 (2020) http://dailynous.com/2020/07/30/philosophers-gpt-3/

(3) Kang, M.: The Mechanical Daughter of René Descartes: the Origin and History of an Intellectual Fable. Modern Intellectual History 14(3), 633-660 (2017)

(4) Cottingham, J.: A Brute to the Brutes? Descartes' Treatment of Animals. Philosophy 53, pp. 551-59 (1978)

(5) Riskin, J.: The Restless Clock: A History of the Centuries-long Argument over What Makes Living Things Tick. University of Chicago Press, Chicago, (2017)

(6) Descartes, R.: Discourse on the Method of Rightly Conducting One's Reason and Seeking Truth in the Sciences in the version translated by Jonathan Bennett, presented at www.earlymoderntexts.com, (2017)

(7) Cavendish, M.: Letter XVIII, Section II. Philosophical Letters. (1664).
Text available at Project Gutenberg,
https://www.gutenberg.org/files/53679/53679-h/53679-h.htm

(8) Turing, A.: Computing Machines and Intelligence. Mind 59(236), 433-460 (1950)

(9) Lin, P., Jenkins R., Abney, K. (eds): Robot Ethics 2.0: from Autonomous Cars to Artificial Intelligence. Oxford University Press, Oxford (2017).

(10) 'Osborne, M., Frey, C.: The Future of Employment: How susceptible are jobs to automation. Oxford Martin School, Oxford, (2013)
https://www.oxfordmartin.ox.ac.uk/downloads/academic/future-of-employment.pdf

(11) Harari, Y.: Homo Deus: A Brief History of Tomorrow. Harvill Secker, London (2015)

(12) Zuboff, S.: The Age of Surveillance Capitalism. Profile, London (2019)

(13) Wang, B., Li, X: Big Data, Platform Economy and Market Competition: A Preliminary Construction of Plan-Oriented Market Economy System in the
Information Era. World Review of Political Economy, 8:2, 138-161, (2017)

(14) Phillips, L., Rozworski, M.: The People's Republic of Walmart: How the World's Biggest Corporations are Laying the Foundation for Socialism. Verso, London (2019)

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

(16) Murray, S.: The Technological Singularity. MIT Press, Cambridge MA (2015)

(17) McCarthy, J.: The Robot and the Baby.
http://jmc.stanford.edu/articles/robotandbaby.html (2004)