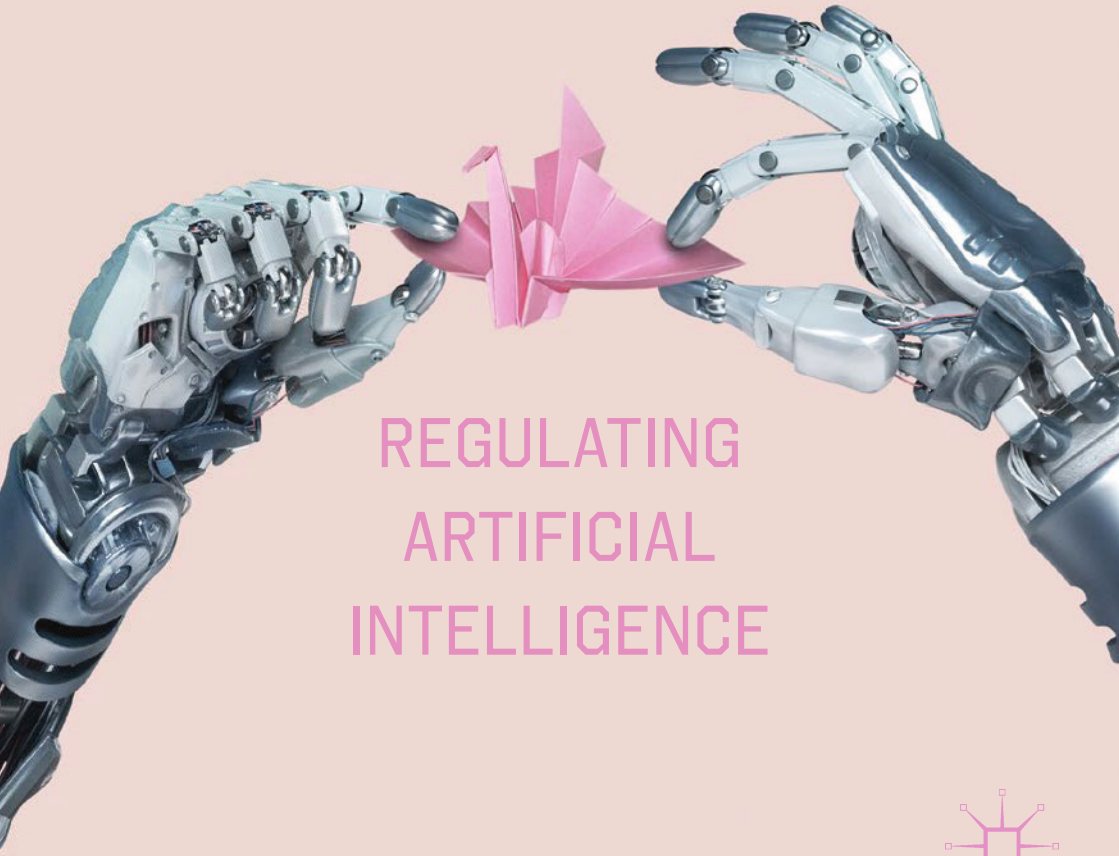
JACOB TURNER

# ROBOT RULES



## REGULATING ARTIFICIAL INTELLIGENCE

# Robot Rules

Jacob Turner

# Robot Rules

## Regulating Artificial Intelligence

palgrave
macmillan

Jacob Turner
Fountain Court Chambers
London, UK

ISBN 978-3-319-96234-4 ISBN 978-3-319-96235-1 (eBook)
https://doi.org/10.1007/978-3-319-96235-1

Library of Congress Control Number: 2018951042



This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

*To Joanne, Caroline and Jonathan*

# Foreword

This is a very timely, thought-provoking and significant book.

These days, even half-serious newspapers contain at least one article every week, sometimes an article almost every day, on some aspect of the imminent and fundamental changes which are (mostly justifiably) said to be about to be wrought to our private, social and working lives by artificial intelligence, or robots. As with many prospective developments, both the precise nature and extent of the changes which AI will cause and the timing of any changes are to a significant extent a matter of conjecture, and so, there is room for a range of respectable predictions. More troublingly, the more extreme, imminent and confident (almost, one can say, the more unrespectable) any prediction about the future, the greater the prominence of the coverage it receives in the popular media. However, there is force in the point that virtually any discussion of the likely effect of a significant prospective development is to be welcomed, as it plays an essential part in the vital exercise of encouraging us to think about and prepare for that development when it comes to pass. Because the potential changes resulting from artificial intelligence will almost certainly be more revolutionary and more widespread than any development since homo sapiens evolved, these factors are all particularly in point when it comes to AI.

Having said that, in a somewhat paradoxical way, the current sensationalist coverage of the likely effects of artificial intelligence seems

almost more to mask, rather than to get people ready for, the extraordinary changes which will result from AI. I think that this is due partly to a sort of novelty inoculation or exhaustion—in other words, the popular media crying wolf too often, too thoughtlessly and too loudly. But an at least equally important factor is, I believe, that the potential effects of AI are so far-reaching in relation to all aspects of our physical, mental, social and moral lives that most people find these changes too challenging to think about in any constructive or practical way. And yet it is both very important and very urgent that we prepare ourselves, both mentally as individuals and structurally as a society, for the AI revolution.

Amidst all the sensationalist, generalised noise, there are a number of much more considered and expert treatments of artificial intelligence, in the form of books and reports. Because the effect of AI is almost certain to be so very far-reaching, there is a need in particular for a considered and informed study of the legal, ethical, and regulatory implications of AI, bearing in mind the many individual areas which are liable to be seriously disrupted, challenged, marginalised or revolutionised as it is rolled out. As Jacob Turner says in this book, the world needs to be as well prepared as it can be for what has been, sensationally if not inaccurately, described as the unstoppable march of the robots—and the sooner we start seriously preparing the better.

A thoughtful and informed book which analyses the implications of current and future developments in AI and how we should plan to deal with them is therefore to be unreservedly welcomed. To write such a book requires a combination of many abilities—including a proper appreciation of the capabilities, functioning, and limits of computer science and technology, a combination of common sense and imagination, an understanding of society, human nature, and economics, and a real appreciation of morality, law and ethics. Not many people have this combination of talents, but any reader of *Robot Rules: Regulating Artificial Intelligence* will, I think, agree that Jacob Turner has demonstrated that he has.

The earlier chapters in this book set the scene and then discuss a number of important and challenging issues of principle and practice which will be thrown up by AI. These chapters include some facts about AI which are not only little known and interesting, but help to explain where we have got to so far. For instance, AI has been with us for well over half a century, in ways which Jacob Turner describes, and this means that we have experience as well as imagination to guide us to the future.

He also explains that AI involves different concepts; indeed, its very definition is a matter of contention, and he provides his own, to my mind rather satisfying, definition.

In addition, when discussing concepts, Jacob Turner brings what could be a dry topic to life by briefly, but illuminatingly, tracing their history and by raising very profound questions. Thus, when considering the question whether robots should have rights, he traces the development of animal rights. And his discussion of the debate as to whether robots can be said to have feelings raises deep metaphysical and moral questions as to the nature of consciousness and compassion, not to mention sex, and even the existence of the human soul. And in the chapter discussing whether robots should have a legal personality, a number of vivid examples are given, including robots in the boardroom and the Random Darknet Shopper.

In two chapters of particular fascination for lawyers, but also for interested non-lawyers, Jacob Turner explains why AI is already starting to require changes to some fundamental legal concepts, such as agency and causation, and he considers how certain principles of liability could be adjusted to incorporate AI—in criminal law, and in negligence, product liability, vicarious liability, contract, insurance and IP in civil law.

There is also an explanation as to how and why AI is an unprecedented technological development, particularly with the advent of unsupervised machine learning—i.e. machines learning without human input (as famously recently occurred with AlphaGo Zero) and no doubt in due course learning from other machines. In a nutshell, it is not only because AI will be so far-reaching in its effects, but also because it will be able to consider issues and resolve them both independently of, and unpredictably to, humans. This gives rise to a host of specific problems, which this book identifies and illuminatingly discusses. In effect, as Jacob Turner suggests, these problems can be divided into three categories, albeit that, at least when it comes to solutions, the three categories will, I think, be interconnected.

First, the issue of rights: should we be granting robots legal personality, like we treat companies, for example, as having a legal personality? To me, the argument that we should do so has real logical attraction. A company cannot act off its own bat: it can only act through humans. AI by contrast, although formed by humans (albeit maybe only indirectly), will be able to act off its own initiative. But the very fact that companies can only act through humans renders the notion of their having legal

personality and liability less threatening to our ideas of normality. Giving robots legal personality brings home to us that, at least in some important respects, they are really like artificial people.

Second, the issue of responsibility: who is liable if AI causes any sort of damage, and who owns the intellectual property which is created by AI? If robots are granted a legal personality, the answer may be simple: the robots themselves. If they are not, then these questions become very thorny, but the answer may lie with their creator or vendor, or, if they are altered or not properly maintained, their operator (if there is such a person). As this book explains, it is probable that issues familiar to lawyers, such as foreseeability and remoteness, will come into play in rather new forms.

Third, ethics: how should AI make choices, and are there any categories of decision which AI should not take? This may well be the most difficult and challenging of the question, particularly if one considers the political and military implications. As Jacob Turner says, the biggest question is how humanity should live alongside AI; some experts believe that the survival of the human race could depend on solving this sort of issue. Further, it is an aspect of AI whose resolution particularly requires worldwide agreement and consistency, and worldwide enforcement that is seen to be effective.

Having raised these questions, Jacob Turner discusses them in a readable and thought-provoking way, which demonstrates that he has studied and thought about the technicalities, principles and practicalities in depth. However, he does not blind or bore the reader with too much or too detail or technicality. He focusses, quite rightly, on both principle and practicality. And, while, very sensibly, he does not suggest that there are any quick and easy answers, he raises and discusses the various options and clearly examines their respective pros and cons.

Having discussed these issues, the book, in an important chapter, containing an interesting review of the current state of play in a number of leading countries, discusses regulation and emphasises the need for global, rather than merely national, rules. Rather than leaving the issue to private groups or companies or to judges, Jacob Turner convincingly opts for legislation and accordingly recommends the development of new public institutions in order to formulate or suggest rules and principles on a cross-discipline and cross-border basis, citing domain names and space law as examples. This book is thus aimed at multidisciplinary audience—from lawyers and politicians to engineers and philosophers,

not only because every thoughtful and responsible person should be interested in this topic, but also because people with all sorts of different expertise and experience will need to contribute to resolving the issues thrown up by AI.

The book then goes on to examine in two chapters the extent and ways in which both the creators of robots and the robots themselves might be controlled, characteristically giving examples of both the provisions of established rules in other fields and how they were actually agreed. And, as Jacob Turner explains, there has already been much work done on these topics in the field of AI itself, and the effect of that work is clearly and trenchantly summarised and assessed. These two chapters, whose contents may sound rather dry, in fact provide a different, and interesting, perspective on the fundamental issues discussed in the earlier chapters.

The book concludes with an Epilogue which in turn ends with the three sentences "In order to write rules for robots, the challenge is clear. The tools are at our disposal. The question is not whether we can, but whether we will". Thanks to Jacob Turner's book, the tools are now more readily at everyone's disposal, and the likelihood of writing the rules and doing so successfully has been substantially increased.

Temple, London EC4
August 2018

David Neuberger
The Rt. Hon. Lord Neuberger
of Abbotsbury, PC, President of
the UK Supreme Court 2012–2017

# Acknowledgements

I have been fortunate to have had the help of many people in writing this book.

The idea for *Robot Rules* came from a lecture which I helped Lord Mance to prepare in 2016, former Deputy President of the UK Supreme Court, when I was his Judicial Assistant (Law Clerk). He was asked to address a conference on the "Future of Law" and the topic we chose was how AI should be regulated. What began as a ten-minute speech has ended two years later as a book spanning several hundred pages.

From the outset, Rob Pilgrim and I shared many stimulating and enjoyable discussions in developing the concept for the book. I owe much to all those who read and provided helpful comments on my drafts, including Oliver Nash, James Tobias, Sean Legassick, Dr. Shahar Avin, Professor José Hernández-Orallo, Matthew Fisher, Nicholas Paisner, Jakob Gleim, Camilla Turner and Gabriel Turner, and several anonymous reviewers.

I have also learned a great deal from exchanges with experts on AI policy including Hal Hodson, Georgia Frances King, Ed Leon Klinger, Dr. Tanya Filer, Guy Cohen, Haydn Bellfield, Rob McCargow, Professor Toby Walsh and Professor Stuart Russell. Jeffrey Ding's insights into China's AI policy were instrumental to the sections on that topic, and Professor Fumio Shimpo provided very helpful information on Japan.

I received valuable feedback on my ideas from the students and researchers on the Masters in Business Technology Course at the Judge Business School, Cambridge University; the Leverhulme Centre for the Future of Intelligence, Cambridge University; Queen's University,

Canada; the Oxford University Computer Science and Law Faculties; and the Future of Humanity Institute, Oxford University. I am very grateful to the conveners of the various courses, seminars and lectures: Assistant Professor Samuel Dahan, Aaron Libbey, Professor Jaideep Prabhu, Neil Gough, Professor Rebecca Williams, Assistant Professor Konatsu Nishigai, Dr. Steven Cave and Dr. Kanta Dihal. In addition, I have found the UK's All Party Parliamentary Group on AI, organized brilliantly by Professor Birgitte Andersen and Niki Iliadis, to be a great source of ideas and debate. I am very grateful to Lord Neuberger, former President of the UK Supreme Court, for kindly agreeing to write the Foreword.

I am indebted to my excellent editorial team at Palgrave Macmillan/Springer, in particular Kyra Saniewski and Rachel Krause Daniel, who have been extremely helpful and informative throughout the commissioning, writing and editing process.

My parents Jonathan and Caroline continue to inspire me, and I am grateful to them for reading and commenting on drafts throughout the process. I would also like to thank my parents in law, Martin and Susan Paisner for all their support and hospitality; it was under their roof that much of this book was written. Finally, I could not have written this book without the love, patience and proof-reading of my wife Joanne.

London, England
August 2018

Jacob Turner

# Contents

CHAPTER 1

# Introduction

> He had not a minute more to lose. He pulled the axe quite out, swung it with both arms, scarcely conscious of himself, and almost without effort, almost mechanically, brought the blunt side down on her head. He seemed not to use his own strength in this. But as soon as he had once brought the axe down, his strength returned to him…. Then he dealt her another and another blow with the blunt side and on the same spot. The blood gushed as from an overturned glass, the body fell back. He stepped back, let it fall, and at once bent over her face; she was dead.[1]
>
> Fyodor Dostoyevsky, *Crime and Punishment*

Our immediate reaction is emotional: anger, horror, disgust. And then reason sets in. A crime has been committed. A punishment must follow.

Now imagine the perpetrator is not a human, but a robot. Does your response change? What if the victim is another robot? How should society, and the legal system, react?

For millennia, laws have ordered society, kept people safe and promoted commerce and prosperity. But until now, laws have only had one subject: humans. The rise of artificial intelligence (AI) presents novel issues for which current legal systems are only partially equipped. Who or what should be liable if an intelligent machine harms a person or property? Is it ever wrong to damage or destroy a robot? Can AI be made to follow any moral rules?

[1] Fyodor Dostoyevsky, *Crime and Punishment*, translated by Constance Garnett (Urbana, IL: Project Gutenberg, 2006), Chapter VII.



J. Turner, *Robot Rules*, https://doi.org/10.1007/978-3-319-96235-1_1

The best-known answers to any of these questions are Isaac Asimov's Laws of Robotics, from 1942:

> First: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
>
> Second: A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
>
> Third: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.
>
> Fourth: A robot may not harm humanity or, by inaction, allow humanity to come to harm.[2]

But Asimov's rules were never meant to serve as a blueprint for humanity's actual interaction with AI. Far from it, they were written as science fiction and were always intended to lead to problems. Asimov himself said: "These laws are sufficiently ambiguous so that I can write story after story in which something strange happens, in which the robots don't behave properly, in which the robots become positively dangerous".[3] Although they are simple and superficially attractive, it is easy to conceive of situations in which Asimov's Laws are inadequate. They do not say what a robot should do if it is given contradictory orders by different humans. Nor do they account for orders which are iniquitous but fall short of requiring a robot to harm humans, such as commanding a

[2] Isaac Asimov, "Runaround", in *I, Robot* (London: HarperVoyager, 2013), 31. Runaround was originally published in *Astounding Science Fiction* (New York: Street & Smith, March 1942). Owing to the potential weaknesses in his first three laws, Asimov later added the Fourth or Zeroth law. See Isaac Asimov, "The Evitable Conflict", *Astounding Science Fiction* (New York: Street & Smith, 1950).

[3] Isaac Asimov, "Interview with Isaac Asimov", interview on Horizon, BBC, 1965, http://www.bbc.co.uk/sn/tvradio/programmes/horizon/broadband/archive/asimov/, accessed 1 June 2018. Asimov made a similar statement in the introduction to his collection *The Rest of Robots:* "[t]here was just enough ambiguity in the Three Laws to provide the conflicts and uncertainties required for new stories, and, to my great relief, it seemed always to be possible to think up a new angle out of the sixty-one words of the Three Laws". Isaac Asimov, *The Rest of Robots* (New York: Doubleday, 1964), 43.

robot to steal. They are hardly a complete code for managing our relationship with AI.

This book provides a roadmap for a new set of regulations, asking not just what the rules should be but—more importantly—who should shape them and how can they be upheld.

There is much fear and confusion surrounding AI and other developments in computing. A lot has already been written on near-term problems including data privacy and technological unemployment.[4] Many writers have also speculated about events in the distant future, such as an AI apocalypse at one extreme,[5] or a time when AI will bring a new age of peace and prosperity, at the other.[6] All these matters are important, but they are not the focus of this book. The discussion here is not about robots taking our jobs, or taking over the world. Our aim is to set out how humanity and AI can coexist.

## 1 Origins of AI

Modern AI research began on a summer programme at Dartmouth College, New Hampshire, in 1956, when a group of academics and students set out to explore how machines could intelligently think.[7]

[4] As to data, see "Data Management and Use: Governance in the 21st Century a Joint Report by the British Academy and the Royal Society", *British Academy and the Royal Society*, June 2017, https://royalsociety.org/~/media/policy/projects/data-governance/data-management-governance.pdf, accessed 1 June 2018. As to unemployment, see Carl Benedikt Frey and Michael A. Osborne, "The Future of Employment: How Susceptible Are Jobs to Computerisation?", *Oxford Martin Programme on the Impacts of Future Technology Working Paper*, September 2013, http://www.oxfordmartin.ox.ac.uk/downloads/academic/future-of-employment.pdf, accessed 1 June 2018. See also Daniel Susskind and Richard Susskind, *The Future of the Professions: How Technology Will Transform the Work of Human Experts* (Oxford: Oxford University Press, 2015).

[5] See Nick Bostrom, *Superintelligence* (Oxford: Oxford University Press, 2014).

[6] See Ray Kurzweil, *The Singularity Is Near: When Humans Transcend Biology* (New York: Viking Press, 2005).

[7] Several nineteenth-century thinkers including Charles Babbage and Ada Lovelace arguably predicted the advent of AI and even prepared designs for machines capable of carrying out intelligent tasks. There is some debate as to whether Babbage actually believed that such a machine was capable of cognition. See, for example, Christopher D. Green, "Charles Babbage, the Analytical Engine, and the Possibility of a 19th-Century Cognitive Science", in *The Transformation of Psychology*, edited by Christopher D. Green, Thomas Teo, and Marlene Shore (Washington, DC: American Psychological Association Press, 2001), 133–152. See also Ada Lovelace, "Notes by the Translator", Reprinted in R.A. Hyman, ed. *Science and Reform: Selected Works of Charles Babbage* (Cambridge: Cambridge University Press, 1989), 267–311.

However, the idea of AI goes back much further.[8] The creation of intelligent beings from inanimate materials can be traced to the very earliest stories known to humanity. Ancient Sumerian creation myths speak of a servant for the Gods being created from clay and blood.[9] In Chinese mythology, the Goddess Nüwa made mankind from the yellow earth.[10] The Judeo-Christian Bible and the Quran have words to similar effect: "And the Lord God formed man of the dust of the ground, and breathed into his nostrils the breath of life; and man became a living soul".[11] In one sense, humans were really the first AI.

In literature and the arts, the idea of technology being used to create sentient assistants for humans or Gods has been around for thousands of years. In Homer's *Iliad*, which dates to around the eighth century BC, Hephaestus the blacksmith is "assisted by servant maids that he had made from gold to look like women".[12] In Eastern European Jewish folklore, there are tales of a rabbi in sixteenth century Prague who created the Golem, a giant human-like figure made from clay, in order to defend his ghetto from anti-Semitic pogroms.[13] In the nineteenth century, Frankenstein's monster brought to the popular imagination the dangers of humans attempting to create or recreate, intelligence through science and technology. In the twentieth century, ever since the term "robot" was popularised by Karel Čapek's screenplay *Rossum's Universal Robots*,[14] there have been many examples of AI in films, television and

[8]What follows is by no means intended to be exhaustive. For a far more comprehensive survey of AI and robotics in popular culture, religion and science, see George Zarkadakis, *In Our Image: Will Artificial Intelligence Save or Destroy Us?* (London: Rider, 2015).

[9]T. Abusch, "Blood in Israel and Mesopotamia", in *Emanuel: Studies in the Hebrew Bible, the Septuagint, and the Dead Sea Scrolls in Honor of Emanuel Tov*, edited by Shalom M. Paul, Robert A. Kraft, Eva Ben-David, Lawrence H. Schiffman, and Weston W. Fields (Leiden, The Netherlands: Brill, 2003), 675–684, especially at 682.

[10]New World Encyclopedia, Entry on Nuwa (quoting Qu Yuan (屈原), book: "Elegies of Chu" (楚辞, or Chuci), Chapter 3: "Asking Heaven" (天問)), http://www.newworldencyclopedia.org/entry/Nuwa, accessed 1 June 2018.

[11]Genesis 2:7, King James Bible.

[12]Homer, *The Iliad*, translated by Herbert Jordan (Oklahoma: University of Oklahoma Press: Norman, 2008), 352.

[13]Eden Dekel and David G. Gurley, "How the Golem Came to Prague", *The Jewish Quarterly Review*, Vol. 103, No. 2 (Spring 2013), 241–258.

[14]The original Czech is "Rossumovi Univerzální Roboti". Roboti translates roughly to "slaves". We will return to this feature in Chapter 4.

other media forms. But now for the first time in human history, these concepts are no longer limited to the pages of books or the imagination of storytellers.

Today, many of our impressions of AI come from science fiction and involve anthropomorphic manifestations that are either friendly or, more usually, unfriendly. These might include the bumbling C-3PO from *Star Wars*, Arnold Schwarzenegger's noble Terminator or the demonic HAL from *2001: A Space Odyssey*.

On the one hand, these humanoid representations of AI constitute a simplified caricature—something to which people can easily relate, but which bears little resemblance to AI technology as it stands. On the other hand, they represent a paradigm which has influenced and shaped AI as successive generations of programmers are inspired to attempt to recreate versions of entities from books, films and other media. In the field of AI, first science then life imitates art. In 2017, Neuralink, a company backed by serial technology entrepreneur Elon Musk, announced that it was developing a "neural lace" interface between human brain tissue and artificial processors.[15] Neural lace is—by Musk's own admission—heavily influenced by the writings of science fiction authors including in particular the *Culture* novels of Iain M. Banks.[16] Technologists have taken inspiration from stories found in faith as well as popular culture: Robert M. Geraci argues that, "[t]o understand robots, we must understand how the history of religion and the history of science have twined around each other, quite often working towards the same ends and quite often influencing another's methods and objectives".[17]

[15] "Homepage", Neuralink Website, https://www.neuralink.com/, accessed 1 June 2018; Chantal Da Silva, "Elon Musk Startup 'to Spend £100m' Linking Human Brains to Computers", *The Independent*, 29 August 2017, http://www.independent.co.uk/news/world/americas/elon-musk-neuralink-brain-computer-startup-a7916891.html, accessed 1 June 2018. For commentary on Neuralink, see Tim Urban's provocative blog post "Neuralink and the Brain's Magical Future", *Wait But Why*, 20 April 2017, https://waitbutwhy.com/2017/04/neuralink.html, accessed 1 June 2018.

[16] Tim Cross, "The Novelist Who Inspired Elon Musk", *1843 Magazine*, 31 March 2017, https://www.1843magazine.com/culture/the-daily/the-novelist-who-inspired-elon-musk, accessed 1 June 2018.

[17] Robert M. Geraci, *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality* (New York: Oxford University Press, 2010), 147.

Although popular culture and religion have helped to shape the development of AI, these portrayals have also given rise to a misleading impression of AI in the minds of many people. The idea of AI as only meaning humanoid robots which look, sound and think like us, is mistaken. Such conceptions of AI make its advent appear to be distant, given that no technology at present comes remotely close to resembling the type of human-level functionality made familiar by science fiction.

The lack of a universal definition for AI means that those attempting to discuss it may end up speaking at cross-purposes. Therefore, before it is possible to demonstrate the spreading influence of AI or the need for legal controls, we must first set out what we mean by this term.

## 2 Narrow and General AI

It is helpful at the outset to distinguish two classifications for AI: narrow and general.[18] Narrow (sometimes referred to as "weak") AI denotes the ability of a system to achieve a certain stipulated goal or set of goals, in a manner or using techniques which qualify as intelligent (the meaning of "intelligence" is addressed below). These limited goals might include natural language processing functions like translation, or navigating through an unfamiliar physical environment. A narrow AI system is suited only to the task for which it is designed. The great majority of AI systems in the world today are closer to this narrow and limited type.

General (or "strong") AI is the ability to achieve an unlimited range of goals, and even to set new goals independently, including in situations of uncertainty or vagueness. This encompasses many of the attributes we think of as intelligence in humans. Indeed, general AI is what we see portrayed in the robots and AI of popular culture discussed above. As yet, general AI approaching the level of human capabilities does not exist and some have even cast doubt on whether it is possible.[19]

[18]For the distinction, see David Weinbaum and Viktoras Veitas, "Open Ended Intelligence: The Individuation of Intelligent Agents", *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 29, No. 2 (2017), 371–396.

[19]See Roger Penrose, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford: Oxford University Press, 1989). The number of sceptics may be shrinking. As Wallach and Allen comment: "pessimists tend to get weeded out of the profession", Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford: Oxford University Press, 2009), 68. For instance, Margaret Boden was one of the most well-known proponents of the sceptical view, although in her latest work, Margaret Boden, *AI: Its nature and Future* (Oxford: Oxford University Press, 2016), 119

Narrow and general AI are not hermetically sealed from each other. They represent different points on a continuum. As AI becomes more advanced, it will move further away from the narrow paradigm and closer to the general one.[20] This trend may be hastened as AI systems learn to upgrade themselves[21] and acquire greater capabilities than those with which they were originally programmed.[22]

## 3 Defining AI

The word "artificial" is relatively uncontroversial. It means something synthetic and which does not occur in nature. The key difficulty is with the word "intelligence", which can describe a range of attributes or abilities. As computer science expert and futurist Jerry Kaplan says, the question "what is artificial intelligence?" is an "easy question to ask and a hard one to answer" because "there's little agreement about what intelligence is".[23]

Some have suggested that the lack of general agreement on a definition of AI is beneficial. The authors of Stanford University's *One Hundred Year Study on Artificial Intelligence* state:

> Curiously, the lack of a precise, universally accepted definition of AI probably has helped the field to grow, blossom, and advance at an ever-accelerating pace. Practitioners, researchers, and developers of AI are instead guided by a rough sense of direction and an imperative to "get on with it".[24]

*et seq* she acknowledges the potential for "real" artificial intelligence, but maintains that "…*no one knows* for sure, whether [technology described as Artificial General Intelligence] could really be intelligent".

[20] See further Chapter 3 at s. 2.1.2.

[21] As to AI systems developing the capacity to self-improve, see further FN 114 below and more generally Chapter 2 at s. 3.2.

[22] Our prediction for the process of narrow AI gradually coming closer to general AI is similar to evolution. *Homo sapiens* did not appear overnight as if by magic. Instead, we developed iteratively through a series of gradual upgrades to our hardware (bodies) and software (minds) on the basis of trial and error experiments, otherwise known as natural selection.

[23] Jerry Kaplan, *Artificial Intelligence: What Everyone Needs to Know* (New York: Oxford University Press, 2016), 1.

[24] Peter Stone et al., "Defining AI", in *"Artificial Intelligence and Life in 2030". One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel* (Stanford, CA: Stanford University, September 2016), http://ai100.stanford.edu/2016-report, accessed 1 June 2018.

Defining AI can resemble chasing the horizon: as soon as you get to where it was, it has moved somewhere into the distance. In the same way, many have observed that AI is the name we give to technological processes which we do not understand.[25] When we have familiarised ourselves with a process, it stops being called AI and becomes just another clever computer programme. This phenomenon is known as the "AI effect".[26]

Rather than asking "what is AI?" it is better to start with the question: "why do we need to define AI at all?" Many books are written on energy, medicine and other general concepts which do not start with a chapter on the definition of these terms.[27] In fact, we go through life with a functional understanding of many abstract notions and ideas without necessarily being able to describe them perfectly. Time, irony and happiness are just a few examples of concepts that most people understand but would find difficult to define. Justice Potter Stewart of the US Supreme Court once said that he could not define hardcore pornography "But I know it when I see it".[28]

However, when considering how to regulate AI, it is not sufficient to follow Justice Stewart. In order for a legal system to function effectively, its subjects must be able to understand the ambit and application of its rules. To this end, legal theorist Lon L. Fuller set out eight formal requirements for a system of law to satisfy certain basic moral norms—principally that humans have an opportunity to engage with them and shape their behaviour accordingly. Fuller's desiderata include requirements that law should be promulgated so that citizens know the standards to which they are being held, and that laws should be understandable.[29] To pass Fuller's tests, legal systems must use specific and workable definitions

[25] Pamela McCorduck, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence* (Natick, MA: A.K. Peters, 2004), 133.

[26] Peter Stone et al., "Defining AI", in *"Artificial Intelligence and Life in 2030". One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel* (Stanford, CA: Stanford University, September 2016), http://ai100.stanford.edu/2016-report, accessed 1 June 2018. See also Pamela McCorduck, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence* (Natick, MA: A.K. Peters, 2004), 204.

[27] The same observation might be made of law itself. See H.L.A. Hart, *The Concept of Law* (2nd edn. Oxford: Clarendon, 1997).

[28] *Jacobellis v. Ohio*, 378 U.S. 184 (1964), 197.

[29] Lon L. Fuller, *The Morality of Law* (New Haven, CT: Yale University Press, 1969).

when describing the conduct and phenomena which are subject to regulation. As Fuller says: "We need to share the anguish of the weary legislative draftsman who at 2:00 a.m. says to himself 'I know this has got to be right, and if it isn't people may be hauled into Court for things we don't mean to cover at all. But for how long must I go on rewriting it?'".[30]

In short, people cannot choose to comply with rules they do not understand. If the law is impossible to know in advance, then its role in guiding action is diminished if not destroyed. Unknown laws become little more than tools of the powerful. They can lead ultimately to the absurd and frightening scenario imagined in Kafka's *The Trial*, where the protagonist is accused, condemned and ultimately executed for a crime which is never explained to him.[31]

Most of the universal definitions of AI that have been suggested to date fall into one of two categories: human-centric and rationalist.[32]

### *3.1 Human-Centric Definitions*

Humanity has named itself *homo sapiens*: "wise man". It is therefore perhaps unsurprising that some of the first attempts at defining intelligence in other entities referred to human characteristics. The most famous

[30] Ibid., 107.

[31] Franz Kafka, *The Trial*, translated by Idris Parry (London: Penguin Modern Classics, 2000).

[32] Stuart Russell and Peter Norvig divide definitions into four categories: (i) thinking like a human: AI systems adopt similar thought processes to human beings;

(ii) acting like a human: AI systems are behaviourally equivalent to human beings;

(iii) thinking rationally: AI systems have goals and reason their way towards achieving those goals;

(iv) acting rationally: AI systems act in a manner that can be described as goal-directed and goal-achieving. Stuart Russell and Peter Norvig, *Artificial Intelligence: International Version: A Modern Approach* (Englewood Cliffs, NJ: Prentice Hall, 2010), para. 1.1 (hereafter "Russell and Norvig, *Artificial Intelligence*"). However, John Searle's "Chinese Room" thought experiment demonstrates the difficulty of distinguishing between acts and thoughts. In short, the Chinese Room experiment suggests that we cannot distinguish between intelligence of Russell and Norvig's types (i) and (ii), or types (iii) and (iv) John R. Searle, "Minds, Brains, and Programs", *Behavioral and Brain Sciences*, Vol. 3, No. 3 (1980), 417–457. Searle's experiment has been met with various numbers of replies and criticisms, which are set out in the entry on The Chinese Room Argument, Stanford Encyclopedia of Philosophy, First published 19 March 2004; substantive revision 9 April 2014, https://plato.stanford.edu/entries/chinese-room/, accessed 1 June 2018.

example of a human-centric definition of AI is known popularly as the "Turing Test".

In a seminal 1950 paper, Alan Turing asked whether machines could think. He suggested an experiment called the "Imitation Game".[33] In the exercise, a human invigilator must try to identify which of the two players is a man pretending to be a woman, using only written questions and answers. Turing proposed a version of the game in which the AI machine takes the place of the man. If the machine is able to succeed in persuading the invigilator not only that it is human but also that it is the female player, then it has demonstrated intelligence.[34] Modern versions of the Imitation Game simplify the task by asking a computer program as well as several human blind control subjects to each hold a five-minute typed conversation with a panel of human judges in a different room. The judges have to decide whether or not the entity with which they are corresponding is a human; if the computer can fool a sufficient proportion of them (a popular competition sets this at just 30%), then it has won.[35]

A major problem with Turing's Imitation Game is that it tests only the ability to mimic a human in typed conversation, and that skilful impersonation does not equate to intelligence.[36] Indeed, in some of the more "successful" tests of programmes designed to succeed in the Imitation Game, the programmers prevailed by creating a computer which exhibited frailties which we tend to associate with humans, such as spelling errors.[37] Another tactic favoured by programmers in modern Turing tests is to use stock humorous responses so as to deflect attention

[33] Alan M. Turing, "Computing Machinery and Intelligence", *Mind: A Quarterly Review of Psychology and Philosophy*, Vol. 59, No. 236 (October 1950), 433–460, 460.

[34] Yuval Harari has offered the interesting explanation that the form of Turing's Imitation Game resulted in part from Turing's own need to suppress his homosexuality, to fool society and the authorities into thinking he was something that he was not. The focus on gender and subterfuge in the first iteration of the test is, perhaps, not accidental. Yuval Harari, *Homo Deus* (London: Harvill Secker, 2016), 120.

[35] See, for example, the website of The Loebner Prize in Artificial Intelligence, http://www.loebner.net/Prizef/loebner-prize.html, accessed 1 June 2018.

[36] José Hernández-Orallo, "Beyond the Turing Test", *Journal of Logic, Language and Information*, Vol. 9, No. 4 (2000), 447–466.

[37] "Turing Test Transcripts Reveal How Chatbot 'Eugene' Duped the Judges", Coventry University, 30 June 2015, http://www.coventry.ac.uk/primary-news/turing-test-transcripts-reveal-how-chatbot-eugene-duped-the-judges/, accessed 1 June 2018.

away from their program's lack of substantive answers to the judges' questions.[38]

To avoid the deficiencies in Turing's test, others have suggested definitions of intelligence which do not rely on the replication of *one* aspect of human behaviour or thought and are instead parasitic on society's vague and shifting notion of what makes humans intelligent. Definitions of this type are often variants of the following: "AI is technology with the ability to perform tasks that would otherwise require human intelligence".[39]

The inventor of the term AI, John McCarthy, has said that there is not yet "a solid definition of intelligence that doesn't depend on relating it to human intelligence".[40] Similarly, futurist Ray Kurzweil wrote in 1992 that the most durable definition of AI is "[t]he art of creating machines that perform functions that require intelligence when performed by people".[41] The main problem with parasitic tests is that they

[38]Various competitions are now held around the world in an attempt to find a 'chatbot', as conversational programs are known, which is able to pass the Imitation Game. In 2014, a chatbot called 'Eugene Goostman', which claimed to be a 13-year-old Ukrainian boy, convinced 33% of the judging panel that he was a human, in a competition held by the University of Reading. Factors which assisted Goostman included that English (the language in which the test was held) was not his first language, his apparent immaturity and answers which were designed to use humour to deflect the attention of the questioner from the accuracy of the response. Unsurprisingly, the world did not herald a new age in AI design. For criticism of the Goostman 'success', see Celeste Biever, "No Skynet: Turing Test 'Success' Isn't All It Seems", *The New Scientist*, 9 June 2014, http://www.newscientist.com/article/dn25692-no-skynet-turing-test-success-isnt-all-it-seems.html, accessed 1 June 2018. The author Ian McDonald offers another objection: "Any AI smart enough to pass a Turing test is smart enough to know to fail it". Ian McDonald, *River of Gods* (London: Simon & Schuster, 2004), 42.

[39]This definition is adapted from that used by the UK Department for Business, Energy and Industrial Strategy, *Industrial Strategy: Building a Britain Fit for the Future* (November 2017), 37, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/664563/industrial-strategy-white-paper-web-ready-version.pdf, accessed 1 June 2018.

[40]"What Is Artificial Intelligence?", Website of John McCarthy, last modified 12 November 2007, http://www-formal.stanford.edu/jmc/whatisai/node1.html, accessed 1 June 2018.

[41]Ray Kurzweil, *The Age of Intelligent Machines* (Cambridge, MA: MIT Press, 1992), Chapter 1.

are circular. Kurzweil admitted that his own definition, "… does not say a great deal beyond the words 'artificial intelligence'".[42]

In 2011, Nevada adopted the following human-centric definition for the purpose of legislation regulating self-driving cars: "the use of computers and related equipment to enable a machine to duplicate or mimic the behavior of human beings".[43] The definition was repealed in 2013 and replaced with a more detailed definition of "autonomous vehicle", which was not tied to human actions at all.[44]

Although it is no longer on the statute books, Nevada's 2011 law remains an instructive example of why human-centric definitions of intelligence are flawed. Like many human-centric approaches, this was both over- and under-inclusive. It was over-inclusive because humans do many things which are not "intelligent". These include getting bored, tired or frustrated, as well as making mistakes such as forgetting to indicate when changing lanes. Furthermore, many cars already have non-AI features which could fall within this definition. For instance, automatic headlights which turn on at night would be mimicking the behaviour of a human being turning the lights on manually, but the behaviour would have been triggered by nothing more complex or mysterious than a light sensor coupled to simple logic gate.[45]

[42] Ibid.

[43] NV Rev Stat § 482A.020 (2011), https://law.justia.com/codes/nevada/2011/chapter-482a/statute-482a.020/, accessed 1 June 2018.

[44] For the new law, see NRS 482A.030. "Autonomous vehicle" now means a motor vehicle that is equipped with autonomous technology (Added to NRS by 2011, 2876; A 2013, 2010). NRS 482A.025 "Autonomous technology" means technology which is installed on a motor vehicle and which has the capability to drive the motor vehicle without the active control or monitoring of a human operator. The term does not include an active safety system or a system for driver assistance, including without limitation, a system to provide electronic blind spot detection, crash avoidance, emergency braking, parking assistance, adaptive cruise control, lane keeping assistance, lane departure warning, or traffic jam and queuing assistance, unless any such system, alone or in combination with any other system, enables the vehicle on which the system is installed to be driven without the active control or monitoring of a human operator (Added to NRS by 2013, 2009). Chapter 482A—Autonomous Vehicles, https://www.leg.state.nv.us/NRS/NRS-482A.html, accessed 1 June 2018.

[45] Ryan Calo, "Nevada Bill Would Pave the Road to Autonomous Cars", *Centre for Internet and Society Blog*, 27 April 2011, http://cyberlaw.stanford.edu/blog/2011/04/nevada-bill-would-pave-road-autonomous-cars, accessed 1 June 2018.