RESEARCH PAPER

Constructivism and its risks in artificial intelligence

Gary R Lea

Private researcher, Queanbeyan, NSW 2620, Australia

ABSTRACT

The research and development of artificial intelligence (AI) technologies involve choices that extend well beyond the search for narrow engineering solutions to problems. The label 'constructivism' is used to capture this larger realm of social choice. Drawing on the history of AI, a distinction is made between limited artificial narrow intelligence (ANI) and artificial general intelligence (AGI). Both forms, the paper argues, carry risks. Following this history, the paper outlines how different approaches to rationality have led to different 'tribes' of AI. No universal model of rationality is available to AI engineers. Choice is everywhere. The paper then moves to an exploration of the links between AI and chess. It argues that chess, far from being an objective measure of rationality and intelligence, reveals the subjective biases and risks involved in the pursuit of AI. The paper moves on to provides examples of various unstable and potentially dangerous race heats taking place in AI, including those among various AI research groups (public and private), among corporations and among states. The final section draws together the various risks of AI.

Introduction

Over the last few years, several prominent science and technology commentators starting with Stephen Hawking and Elon Musk (Holley, 2016; D'Orazio, 2014), together with some experts in the field, such as Stuart Russell (Bohannon, 2015), have argued that artificial intelligence (AI) is a technological development of transformational capacity, classifying it as 'high-risk' – alongside nuclear technology. This paper focuses on the constructivist possibilities of AI and its risks.

Risk itself is a fluid and contested concept, viewed by some as an objective given arising out of physical facts (Krebs, 2011) and by others as a subjective given grounded in social construction (Wynne, 1992); following Hansson (2010), this paper starts from the position that it is a complex amalgam of both. AI is clearly something different from any philosophical notion of 'natural kind' (Ellis, 2001; Bird, 2010). The words 'artificial' and 'intelligence' are not merely fuzzy but are signifiers of things that are heavily the product of choices involving both theory and values. The label 'constructivism' is used here to capture the idea that engineering choices in AI are deeply bounded by, and infused with, theories – theories of intelligence, of thinking, of rationality, of human nature and ultimately of the human in nature. Indeed, whether always explicitly recognised as such, it is the very potential of such constructivism in AI that has seen growing calls, including calls from some working in the field, for urgent research not only into technical aspects of AI safety and machine ethics, but also into the social and economic impacts of AI (Russell *et al.*, 2015).

A further complication is that, despite the pervasiveness of the AI tag and associated memes, there is no single understanding of AI, there is no unified approach to AI technology, there is no one way to design or build AI systems. Instead, at least for the present, there is something more akin to

ACCEPTING EDITOR: Peter Drahos

a jungle ecosystem in which various techniques and technologies (relabelled as AI) play various roles in various fields of application analogous to ecological niches.

The four sections of this paper examine the history and development of AI; the creeping dominance of a rational agent approach to AI; the use of the game of chess (a construct) to measure the progress of AI; and the use of AI in familiar social and economic races, such as those among companies or among nations. The upshot is that AI's constructivism is shown to be pervasive. Further and crucially, it transpires that many of the risks of this constructivism turn out to be familiar because they reproduce what we have always done.

A brief history of AI

Although work on various mechanisms imitating humans goes back at least two millennia (Nilsson, 2010), it was developments in electronics before and during World War II that allowed for effective laboratory experimentation: in 1943, Warren McCulloch and Walter Pitts constructed an electronic 'neural element' to provide a simple simulation of a human neural cell (McCulloch and Pitts, 1943). In 1948, Donald Hebb discovered the electrical networking mechanism that allowed human neurons to engage in associative learning, something that was later transferred into hardware (Hebb, 1949; Rosenblatt, 1958. Meanwhile in Britain, Alan Turing wrote about the possibility of developing machine intelligence through computational means, including his now-famous 'Turing test' paper of 1950 (Turing, 1950).

From an Anglophone perspective, one significant point in time was August 1955 when a young US researcher called John McCarthy put forward a proposal for funding a ten- person, two-month workshop-based study of what he called 'artificial intelligence', the first use of the term:

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. (McCarthy *et al.*, 1955)

Funding was granted and the Dartmouth College summer research project on artificial intelligence went ahead in 1956. Now better known as the Dartmouth conference, leading first-wave US-based researchers, such as Marvin Minsky, Claude Shannon, Ray Solomonoff, Arthur Samuel, Herbert Simon and Allen Newell, all attended, laying the foundations for the modern discipline. Parallel but largely separate developments also occurred behind the Iron Curtain under the rubric of Soviet cybernetics (Golubev, 2014.

Following fair progress in the 1950s and 1960s, at least in the West, AI entered a cycle of boom and bust, the bust periods being the first two so-called 'AI winters' (Crevier, 1993). The first AI winter of 1974 to 1980 was triggered by the funding cuts of the defense advanced research projects agency (DARPA) in the US and the Lighthill report to the UK government recommending funding withdrawal. The situation stabilised briefly before the second AI winter of 1987 to 1993, triggered by the termination of the US strategic computing initiative, the collapse of the Japanese fifth generation computer initiative funded by the Japanese ministry of trade and industry (MITI) and the commercial failure of expert systems of that time. Accordingly, from the mid-1990s, most researchers focused on more specific areas of research, such as speech recognition, natural language processing and visual pattern recognition.

Importantly, this narrowing of focus led to what is now called 'artificial narrow intelligence' (ANI), systems with limited capacity in a specific domain of activity. In its ANI guise, AI has been with us since at least the turn of the millennium, and is now at work in our smartphones, our home automation devices and so on. However, ANI is constantly becoming more sophisticated in that both its capacities and domains are broadening, including recombination with robotics. An example of this is the interaction and use of multiple sensor inputs and control systems in autonomous vehicle systems (Sjafrie, 2019).

Fruitful progress with ANI has still left open the question of whether artificial general intelligence (AGI), a system with human or greater competence across human or greater domains, was achievable. One key problem was defining and benchmarking what intelligence in relation to AGI actually was. John McCarthy spelt out the problem:

> Q. What is artificial intelligence?
>
> A. It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.
>
> Q. Yes, but what is intelligence?
>
> A. Intelligence is the computational part of the ability to achieve goals in the world. Varying kinds and degrees of intelligence occur in people, many animals and some machines.
>
> Q. Isn't there a solid definition of intelligence that doesn't depend on relating it to human intelligence?
>
> A. Not yet. The problem is that we cannot yet characterise in general what kinds of computational procedures we want to call intelligent. We understand some of the mechanisms of intelligence and not others. (McCarthy, 2007)

However, shortly after McCarthy wrote this, two significant events occurred. First, what was now the AGI research community felt that sufficient progress had been made to initiate a series of global conferences, the first one held at the University of Memphis in 2008 (de Garis and Goertzel, 2009). Second, Marcus Hutter (now of the Australian National University) and Shane Legg (now of Google DeepMind) sought to define intelligence in formal mathematical terms for the first time; from their ground-breaking mathematical formalisation, known as AIXI, an informal Legg-Hutter definition emerged:

> Intelligence is the measure of an agent's ability to achieve goals in complex environments. (Legg and Hutter, 2007)

It is important to emphasise here that this definition does not seek to cover all of what might typically be regarded as intelligence in humans. This is because the main focus of AI research is usually on modelling certain aspects of human rationality rather than on the full spread of human cognition, much less mental and related activity, such as emotions (Simon, 1967). Although some might argue with the Legg-Hutter definition, it is nevertheless a significant point of reference because, being increasingly adopted (whether on purpose or by default) within AI R&D, it influences what is designed and built.

At this point, it should be emphasised that, although ANI is becoming more sophisticated, it does not automatically follow that AGI will ever be achieved; indeed, surveys of AI experts over the last few years (Barrat, 2013; Müller and Bostrom, 2014; Azulay, 2019) find that at least 10% of those responding thinking that it was unlikely that AGI would ever be achieved, with the possibility of many more 'naysayers' not responding (Fjelland, 2020). The majority opinion from the surveys is that:

1) ANI technology will continue to improve and the rate of technological improvement will accelerate;
2) at least some elements of ANI technology will both broaden out and scale up, indirectly contributing to moves towards AGI;

3) from whatever foundations it comes, there will be further developments and improvements to the point where AGI is eventually achievable; and
4) AGI is more likely than not to be achieved by 2060, and perhaps as early as the 2020s.

Noting the many subsets of each and even hybrids, there are three general foundational approaches to building AI – symbolic-computational, connectional and neuromimetic. The symbolic-computational approach – what used to be called 'good old fashioned AI' (GOFAI) – typically looks past the human brain, seeking via algorithmic means to model and implement a rational agent, a counterpart to the rational actor in the economic and corresponding legal theory literature (Finkelstein, 2004). The agent will typically be performing something like expected utility maximisation or satisficing in the human context (see Schwartz *et al.*, 2002). However, for the agent to learn from its environment and seek to achieve goals in it, a sophisticated cycle of observation, learning, prediction, planning, decision, action and reward are required. In AI systems more explicitly based on AIXI research, implementing this cycle involves mathematical formalisations based on Ockham's razor (Grunwald, 2007), Epicurus' principle (Hutter, 2009) and Kolmogorov complexity (Li and Vitanyi, 2014).

Next, the connectional approach; this takes inspiration from the human brain in the shape of artificial neurons connecting in artificial neural networks. Since Hebbs and McCullough's early experiments, the sophistication of artificial neural network design has expanded considerably (Hagan *et al.*, 2014). In recent years, there has been particular emphasis on so-called 'deep learning'. In a deep learning neural network, raw data in datasets are transformed through layers of artificial neurons into representations which the network can then learn from, often semi-supervised or even unsupervised (Schmidhuber, 2015).

Finally, the neuromimetic approach; sometimes treated as a subset of the connectional approach, looks at the human brain directly and attempts to copy parts or the whole of it, including whole brain simulation. Although intended for medical experiments rather than an AI system, the blue brain project, headed by Henry Markram at the École polytechnique fédérale de Lausanne (EPFL), sought to produce a fully functioning simulation of a human brain down to the molecular level by 2023. After being absorbed by the much larger human brain project, Markram's work program became bogged down in scientific, political and funding controversies (Yong, 2019). Despite this, technical progress was made in a number of areas (Markram *et al.*, 2015).

Translating philosophies of mind into AI: assumptions underlying approaches

When stripped of its detailed mathematics and logic, AI involves the translation of philosophies of mind – here in the broadest sense including the underlying hardware of the brain – into science and engineering (Mandik, 2013). Whether or not always recognised by AI practitioners, this exercise in translation is confronted by different philosophical assumptions about mind and the limitations of these assumptions. Consider, in this context, the work of a quartet of Nobel laureates: Becker, Hayek, Simon and Kahneman. Their peculiar relevance lies in mapping their spectrum of views to the choice of AI research approaches noted by Norvig and Russell (2020): acting humanly (the Turing test approach), thinking humanly (the cognitive modelling approach), thinking rationally (the laws of thought approach) and acting rationally (the rational agent approach). Tersely put, the exploration of economic rationality reveals differences of approach and conception. A global model remains elusive.

Norvig and Russell point out that, although all four research approaches have been visible in AI research, the dominant approach is now the rational agent approach; this is characterised by focus on external measurables against an idealised performance measure (rationality). The rational agent approach has become preferred because there are more avenues to rationality than correct inference under the laws of thought approach and it is "more amenable to scientific development than are approaches based on human behaviour or human thought" (Norvig and Russell, 2020, p.4). Thus, the resulting rational agent:

> [will] operate autonomously, perceive [its] environment, persist over a prolonged time period, adapt to change, and create and pursue goals . . . [it] acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome. (Norvig and Russell, 2020, p.5)

The rational agent approach directly links to rational choice theory (RCT) in economics and the economic approach to human behaviour in sociology (Becker, 1976). At its base, RCT posits that, at least at the behavioural level, individuals act to maximize utility and, correspondingly, behaviour in a society is simply the aggregate of such individual choices (Elster, 1989).

However, even if RCT can provideadequate explanation (Becker, 1957) and even prediction in some very specific settings (e.g., criminal appeals decision-making: Songer *et al.*, 1995), we know that it fails on both counts in many others. For example, Becker's RCT-based arguments for longer sentences as a means to reduce crime rates were simply not supported by the empirical evidence (Nagin, 2013). More nuanced approaches, including those of the other laureates, would hold this is because, even at the behavioural level, criminals, like other humans, operate neither perfectly consistently nor perfectly rationally. In the AI context, this fundamental mismatch between the RCT model of *homo economicus* and actual humans strongly suggests that, if the rational actor paradigm continues to be pursued in AI R&D, the mere fact that the resulting artificial agents 'act rationally' – instead of 'thinking humanly' or at least 'acting humanly' – will generate just those conflicts of goals, values and beliefs that will create significant risk over time for all humanity (Bostrom, 2014). Put another way, in the absence of effective goal constraint or other safety measures, more powerful rational agents could be inherently unsafe with, or for, humans.

A more subtle view was presented by Frederick Hayek, whose work is relevant in connection with uncertainty and AI. In rejecting notions of central economic planning in favour of market pricing as the key mechanism of economic coordination, Hayek highlighted what would today be called 'incomplete information' in terms of the scattering of knowledge among individuals and the importance of some contextually specific (local) knowledge over general knowledge (Hayek, 1945). He also accepted what would now be called 'bounded rationality', in that he rejected radical behaviourism and the import of natural sciences approaches into economics. He acknowledged complexity in society, and accepted the significance of customs, conventions and tacit knowledge (Oguz, 2010). This is linked to what was then a radical connectional model of brain/mind under which there is separation of the physical and the phenomenal aspects of sensation and perception with each classification also being an act of interpretation (Hayek, 1952). Connectional AI systems in the form of neural networks are now commonplace and Hayek's philosophical contribution to the philosophy of AI in this area is increasingly recognised (Steele, 2002). Thus, remarkably to some perhaps, Hayek's work is more situated in 'thinking humanly'.

Although agreeing with Hayek on incomplete information and bounded rationality in general terms, Simon developed a more formalised approach to the latter, replacing the notion of utility maximisation with satisficing (reaching an acceptable level set against aspiration) (Simon, 1957). Furthermore, although he accepted that market pricing was important, Simon felt this was limited to known or reasonably predictable pricing: where prices became unpredictable, other means for coordination and problem-solving were needed and these came from social identification within professional, ethnic and other groups. Significantly, Simon's work on bounded rationality came from his own work within the AI field itself (Simon, 1996), but, unlike Hayek's connectionism, was grounded in algorithmic AI. Nevertheless, for Simon, 'acting humanly' was important; this would explain his argument for states that are analogous to emotion in AI systems.

Finally, Kahneman examined cognitive biases and heuristics operating against standard assumptions of rationality in economics; for example, his and Tversky's prospect theory (Kahneman and Tverksy, 1979) provided a model of decision making that envisaged losses hurting more than gains, and people giving high probabilities greater weight than low probabilities. Subsequent work in the psychology of risk (Kahneman *et al.*, 1982), expertise and expert performance (Kahneman and Klein, 2009), and his contribution to 'nudge' theory (Thaler and Sunstein, 2008, a significant (if somewhat controversial) governance instrument, are highly relevant and examples of the 'acting humanly' approach.

So, we see that the translation of mind into the AI engineering schema requires the engineer to confront the limits of rationality, to understand the extent to which the strength of that rationality is strengthened by its participation in social networks and, ultimately, to make choices about the very direction and purpose of the schema. Unfortunately, this requirement is seldom if ever met, on account of resource limitations, in-built distortions and biases of the translation process itself and the ongoing contest for dominance among schools of thought perceptively referred to by Domingos (2015) as 'tribes', summarised for convenience in Table 1.

The chess problem or the problem with chess? Models and measurement

From philosophies of mind, we turn next to chess. Chess provides a nice example of constructivism in action when it comes to the engineering of AI. At least until 1997 when Deep Blue beat international grandmaster Garry Kasparov, chess was frequently cited by AI experts as the *drosophila* of AI research (Ensmenger, 2011). This is a reference to the common fruit fly, *Drosophila melanogaster*, a so-called 'model organism' for biological research for over a century, including, most recently, genetics and developmental biology (Hales *et al.*, 2015).

This section draws on personal experience of designing and coding a chess program. What does the ability of a machine tell us either about intelligence in general or the nature of human or machine intelligence more specifically? The answer may be very little and certainly very different

Table 1. The tribes of AI

Tribe	Belief	Problem	Solution	Tribal algorithm
Symbolists	Intelligence = symbol manipulation	No matter how intelligent, can't learn from scratch: seed knowledge needed	Identification of missing knowledge to allow deduction to go through and make as general as possible	Inverse deduction
Connectionists	Learning = brain function: reverse engineer from neurons. Brain learns by adjusting strength of neural network connections	Errors can occur in neurons and propagate	Identification of neurons in error and changing	Backpropagation
Evolutionaries	Natural selection = learning. Natural selection can be simulated	Learning structure can go awry and need complete readjustment	Mating and evolving programs to update learning structures	Genetic programming
Bayesians	All learned knowledge = uncertain. Learning = uncertain inference	Dealing with noisy, incomplete and contradictory information	Apply probabilistic inference	Bayes' theorem and variants
Analogizers	Learning = recognising similarities between situations and inferring other similarities	Judging how similar	Determining which experiences to remember and combining for new predictions	Support vector machines (SVM)

from what some technovangelists might have us believe. Having been taught the basics of the modern standard game by my maternal grandfather during the 1970s, my first encounter with AI was through chess programs in the early 1980s. In 1982, I acquired my first computer, a £69.95 Sinclair ZX81. One of the reasons for its low cost was the miniscule 1024 bytes (1KB) of random-access memory (RAM) in the supplied system. Screen display data could consume up to 793 bytes and system variables another 125 bytes, meaning that, unless one could afford a separate RAM expansion board, ZX81 programs had to be coded both compactly and efficiently, and written largely or wholly in Z80 microprocessor machine code (Baker, 1982).

Between December 1982 and February 1983, a three-part paper appeared in *Your Computer* in which David Horne provided listings for an all-machine code chess program that would run on the unexpanded ZX81. Although it was subject to a number of gameplay limitations (restricted start positions and no implementation of castling, promotion or *en passant*), Horne's *1K ZX Chess* operated in a mere 672 bytes. I attempted to write a smaller, better program but, after three months, had something subject to the same limitations, only marginally better in play and needing more RAM, leading to occasional 'out of memory' crashes. For the record, Horne's code compactness record on ZX81 hardware was not surpassed until January 2015 (Ulanoff, 2015) and the 1K ZX81, the first program fully compliant with the rules of the international chess federation (FIDE), was not released until February 2016.

At the time, in common with many others (Rasskin-Gutman, 2009), I believed that the human ability to play chess required high intelligence in the (limited) sense of rational problem-solving ability. On that basis, to the extent a computer could play chess at all, I considered it displayed some intelligence and, the better it played, the more intelligent it was. Furthermore, the compactness and efficiency of Horne's code hinted at the spine-tingling possibility of great leaps in intelligence on better hardware. I was totally wrong, of course, because, at that time, I failed to appreciate the limitations of a narrow and mechanistic view of intelligence, both generally and in the specific context of chess. However, to understand better those limitations, we turn to chess among humans – h2h chess, for convenience.

H2h chess is dominated by males: as at the start of September 2020, the top 84 ranked FIDE chess players in the world all identify as men. In April 2015, when asked to comment on this sort of disparity, British grandmaster Nigel Short infamously asserted that:

> [g]irls [*sic*] just don't have the brains to play chess . . . [they] are hard-wired very differently. Why should . . . [men and women] function in the same way? I don't have the slightest problem in acknowledging that my wife . . . possesses a much higher degree of emotional intelligence than I do. (Short quoted in Watson, 2015)

In response, Judit Polgár, a leading Hungarian player who had beaten Short on a number of occasions, argued that, rather than some bogus lack of rational problem-solving capacity, the real issue was sexism in the game from junior entry onwards and the greater use by males of 'psyching' tactics, including deliberately intimidating body language, also heavily influencing outcomes (cf. Rayman, 2015). Unsurprisingly, there was further evidence in support of Polgár's argument, including females being driven out of early development programmes by males (Galitis, 2002; Blanch, 2016), statistically significant effects arising from the consequent massive reduction in the pool of female players (Bilalić *et al.*, 2009) and experiments showing significant increases in measured performance when females played remotely against males (Backus *et al.*, 2016).

Thus, h2h chess, especially at higher levels, makes rational intelligence a necessary but not sufficient condition to win and even then only with numerous caveats. Although intelligence (as indicated by IQ) is a reasonable predictor of chess-playing ability in novices (De Bruin *et al.*, 2014), the number of practice hours and ability to recall positions and suitable moves from them (i.e., the development of a dynamic, in-memory playbook) become progressively much stronger predictors at each level beyond novice (Bilalić *et al.*, 2009). This seems to link directly to the way humans actually play chess: in order to play with limited resources within a possible 'game space' of a minimum 10^120 properly playable games (the so-called 'Shannon number', calculated by Claude Shannon, the information theory and AI pioneer: Shannon, 1950), human chess players appear to operate combined move search and evaluation, but at a relatively shallow depth, typically observed at fewer than ten half-moves (ply) ahead. However, the tight coupling of this search and evaluation to the developed playbook then allows conception of a broader range of alternative positions (around 40 to 50 for a grand master) and pursuit of broader, dynamically changeable goals, including developing rearrangeable suites of moves towards an end game.

In contrast, the vast majority of modern chess engines go far deeper in a separated first search phase than humans do – as far out as 60 ply – and, treating the sets of possible moves as an inverted 'tree' to be searched, typically use an algorithm called 'alpha-beta deep-first' to 'prune the tree branches' in accordance with move comparison scoring. This exercise in computational brute force is typically followed by a distinct but similarly brutal evaluation of the remaining branches, starting from the summation of assigned piece values. Although most modern chess engines also use databases of openings and endings as a boost to performance, their operation is typically neither as dynamic nor as well-integrated in search and evaluation as the playbook of their human counterparts.

Even if we leave aside the non-rational intelligence aspects of h2h chess, it is clear that operations within most chess engines bear only the most superficial resemblance to the mental processes humans utilise to play chess. This, in turn, leads to a fundamental point about many AI technological artefacts: whatever the historical intentions of such pioneers as Clause Shannon, modern AI does not emulate human intelligence. Instead, it typically and by default simulates some limited aspects of intelligent behaviour. Just as importantly, the chess example reveals that there are more forks in the road when it comes to the development of AI, forks that could lead to different kinds of problem-solving capacities, ones less reliant on computational brute force.

AI in social and economic races

Over the last few years, AI has become the focus of several heats within the context of a global AI race (Prakash, 2017; Walch, 2020). These race heats include those among various AI research groups (public and private), among corporations and among states. Starting in the Asia-Pacific region, for example, while always watching what the current race leader – the US – is doing and wondering what North Korea is up to (Lim, 2019), South Korea, China and Japan regard themselves as competitors in developing and implementing AI and allied technologies not only for commercial and civilian applications (Lo, 2017), but also for military and security ones (Jha, 2016). Likewise, in the US itself, moving beyond existing remote-piloted drones, plans for future battlefield and covert operations include humans and autonomous land, sea and air robots working side by side (Hambling, 2018). Next generation cybersecurity systems powered by more advanced AI technologies will provide not only enhanced defence and national security measures, but also offensive capabilities. From battlefield logistics and medical support to the next generation of pilotless planes, AI will be utilised (Zeimpekis *et al.*, 2014). How to implement the laws of war in code (Tzafestas, 2016), recognise surrender (Sparrow, 2015) and integrate machines into a human military command hierarchy are problems that arise – always with concerns of getting it wrong and losing control (Rasch *et al.*, 2003).

However, states and governments are not interested only in military prowess or the national security aspects of AI. Realising the possible scientific, technological and economic gains that might be made or the losses that might accrue if others are ahead, massive public investment programs have started in Japan, South Korea and China. In addition to existing expenditure via university and corporate R&D schemes, in less than one year to September 2016, Japan committed some $US974 million on a ten-year AGI flagship project at its new Riken Center for advanced integrated intelligence (Demura, 2016. Within days of Google DeepMind's AlphaGo system beating

South Korea's top go player in March 2016, South Korea announced an additional public expenditure of $US890 million on AGI R&D, coupled with private contributions of another $US2.23 billion up to 2020 (Basu, 2016). China's initial response was to make AI a national priority, accounting for a significant portion of massively increased science and technology spending in the following five-year plan period (McClaughlin, 2016). In consequence, the US became increasingly concerned about Chinese actions (Mazur and Markoff, 2017), and reprioritised AI R&D in February 2019 by executive order and subsequent funding arrangements;(White House Office of Science and Technology Policy, 2020). China, meanwhile, has long-since publicly stated its ambitions to lead the world in AI (Jing, 2017).

The number of areas in which AI technologies already have civilian and commercial applications has expanded rapidly. Most obviously, there are developments in cancer diagnostics (Leachman and Merlino, 2017), portfolio management and insurance underwriting (McCurry, 2017), financial news and sports reporting (Simonite, 2015), plus combined adaptive cruise control, lane changing and auto-navigation in cars (Condliffe, 2016 with fully autonomous vehicles to follow from 2021. Having acquired a bad reputation for endless menu sequences and unusable querying, existing 'smart' systems are being replaced or upgraded. New AI-based telephone assistants can provide fully contextualised advice and call placement, including responding to customer emotions detected by real-time analysis of pitch, tone and verbal content (Brewster, 2016).

AI-based trading platforms have already shown consistently better returns than their human counterparts. It is sobering to realise that, in 2000, Goldman Sachs employed some 600 equity traders on Wall Street while, by 2017, it employed just two (for regulatory sign-off purposes). Some 200 AI system engineering jobs were created at the firm, significantly lower paid (Byrnes, 2017). Entire financial markets have disappeared into server farms, including the New York mercantile exchange (NYMEX), which, after reluctantly allowing its first electronic futures trade in late 2006, closed its entire trading floor in December 2016. Craig Weinstein, formerly a NYMEX pit trader earning over a million dollars annually, turned to selling golf green fertiliser in Arizona:

> I'm down to probably my last 50 grand at this point . . . one guy put a gun to his head and killed himself. It's pretty amazing what technology has done to that market. (Weinstein quoted in Meyer, 2016)

Few will be inclined to feel sympathy for Wall Street traders, but these sorts of figures and the fall of a powerful financial elite suggest there will be no smooth passage to jobs in other segments of the global economy. Even AI engineers should not feel safe; some types of AI software are already writing themselves (Simonite, 2017).

The corporate environment in which these changes have taken place is itself highly dynamic. Google, Apple, Facebook, Microsoft, IBM, Amazon and other large, well-established high technology corporations began to compete with each other by providing an array of AI-based products and services (Markoff and Lohr, 2016). In voice query alone, there have been Google Voice Services, Siri, M, Cortana, Watson Voice and Echo. These big beasts also compete for graduates in computer science, maths and engineering, driving up salaries for AI and robotics talent to unprecedented levels (Tilley, 2017). Meanwhile, they must monitor not only their peers, but also the AI SMEs and AI startups. The question is always whether the threatening and/or useful should be bought. DeepMind, the race leader in deep learning, was purchased by Google for over $US500 million despite being less than five years old (Gibbs, 2014). Meanwhile, China's large high technology corporations, such as Baidu, Tencent and Alibaba, have founded and funded AI labs at levels comparable to their western counterparts, frequently luring talent away from them (Perez, 2017; Yang, 2017).

On occasion, as with China's Ant Financial, competition has arisen from organic corporate evolution (Knight, 2017). More usual is a 'dash for cash'. Although superficially autonomous vehicles represent a natural development from ride-sharing, Uber's involvement actually came about through a deliberate crash development program run by engineers previously employed by the

Alphabet driverless car unit, Waymo (Bhuiyan, 2017). Uber found itself embroiled in a lawsuit around misappropriation of trade secrets by its lead project engineer, a person previously employed by Waymo and later convicted in trade secrets criminal proceedings (Korosec and Harris, 2020). Yet, alongside the competition, there can be cooperation: standards setting and research into machine ethics are being conducted jointly by some of the larger US corporations (Horvitz and Suleyman, 2016) while initiatives to create open source AI tools and safety protocols have been started (e.g. Open AI).

Universities and public research institutes in the UK, US, Canada and the EU find themselves hard-pressed. Although their sponsors – public and private – frequently offer extra funding, these institutions complain that their brightest and best graduates go straight to industry (Kunze, 2019). Many senior academics also engage in significant consulting or corporate activity (*Economist*, 2016). AI scientists and engineers in public institutions want at least peer recognition; but fame and fortune beckon for many young PhD graduates lured directly into industry. By 2016, Chinese universities and research institutes had caught up with their US counterparts in terms of AI papers presented at top-ranked conferences and papers on deep learning in top-ranked journals (Fung, 2016). The international association for the advancement of AI annual conference for 2017 had to be rescheduled at the last minute because it clashed with Chinese new year and over 40% of papers submitted were from authors based in China (Zhang, 2017).

What can be learnt from following the money? Certainly, large US-based corporations have ramped up spending on AI and related projects relative to other areas of their business over the last few years. In 2015/2016, the Alphabet group alone was spending in excess of $US3.5 billion in speculative R&D, much of it on AI and related areas (Popper, 2016). IBM has invested heavily in its Watson division, covering not only the natural language query system that made its initial reputation by beating human players at jeopardy! but a whole raft of so-called 'cognitive' services and tools from speech-to-text through to data analytics for large *corpora* of unstructured data (Waters, 2016b). When faced with real-world challenges in healthcare – in particular, oncology (Strickland, 2019) – some Watson services proved to be less robust and cost-efficient than hoped. However, there have been corporate successes elsewhere: IBM's TrueNorth neuroprocessors – microprocessors closely modelled on human brain synaptics – have been acquired for work on next generation drone development by the USAF under the Blue Raven program (Wolfe, 2020). Meanwhile, Microsoft, Facebook and Google have established several research groups, some (semi-) secret for a variety of commercial, military and national security reasons, and have been spending large sums (Bright, 2016; Levy, 2017). Having acquired the Nervana start-up for $US400 million in 2016, Intel internally reorganized itself to create a whole new artificial intelligence products group headed by Nervana's former CEO (Moorhead, 2017).

Perhaps more interesting, though, is what has happened with international corporate investments and international venture capital (VC) flows. Although military and diplomatic tensions between China and the US began to rise significantly after 2014 with hacking allegations and the South China Sea disputes (Smith, 2017), the offshore arms of American VC funds remained among the largest suppliers of AI foreign investment into mainland China (Soo, 2017). Conversely, out of a national overseas technology spend by China in excess of $US225 billion during 2016, Chinese companies invested an estimated $US12.4 billion in the period January 2015 to September 2016 in Silicon Valley startups alone, frequently focusing on AI and robotic systems with military and security applications (Mazur and Perlez, 2017). Yet, with all the subsequent high technology foreign investment curbs, forced divestments (Liao, 2020), product/service bans (Ruehl, 2020), pressure from military and national security organisations (Dumont, 2019) and other actions or threats (Dent, 2020), the Sino-US AI VC pipelines still operate (Fannin, 2020) and the larger companies in each country still retain significant established connections with their counterparts. These include Google and Microsoft Labs in Beijing, and Baidu Labs in Silicon Valley and Seattle. VC and related capital flow tracing also reveals other items of interest: for example, a country that seldom registered on the radar of AI news reporting – Israel – contains some of the most significant AI R&D hubs and has become one of the largest recipients of international AI VC funding (Behar, 2016; Zeff, 2016).

Despite occasional fears in some quarters of yet another AI hype-bubble-burst and a subsequent 'winter', the bandwagon of AI research, development and commercialisation has shown little signs of slowing. On the supply side, even established technology companies have seen AI and robotics as the next growth areas as what were previously profitable specializations in hardware, software or services become fully commoditised or obsolete (Inagaki, 2016; Dwoskin, 2017). On the demand side, corporate adopters of AI and robotics see the possibilities for slashing costs, greater efficiencies and beating competitors at a time when, at least in developed economies, an ageing workforce, stagnant productivity and declining returns from previous rounds of ICT innovation (Adler, 2017) might have proved insuperable barriers.

Taken together, the various AI race heats all contribute to an immensely potent and unstable technological dash with combined aspects of the space race, the arms race and the gold rush, a hurdles event with high stakes and the possibility of injury for participants. It must be remembered, though, that the word 'participants', here, potentially includes nearly all inhabitants of developed, and many inhabitants of developing, countries; at the very least, we can already see that consumers of social media services are already living inside a series of worldwide social experiments with *de minimis* substantive controls or ethics protocols.

The risks of AI reconsidered

For present purposes, let us treat risk simply and basically as 'the possibility of a negative outcome' (Brunschot and Kennedy, 2008) and 'governance' as 'the management of the course of events inside asocial system' (Burris *et al.*, 2005). Derivatively, then, risk governance becomes management of the course of events pertaining to risk within a social system. Given what has been shown above, what does risk governance involve in the AI context?

The one aspect of risk most commonly considered in relation to AI is existential risk envisaging a possible future in which one or more AI systems end up destroying humankind.In the popular press, this is portrayed as the uprising of the machines leading to a doomsday scenario (Madrigal, 2015). Discussion of this existential risk is sometimes linked to the notion of the AI singularity (Good, 1962), the hyper-acceleration of AI development via recursive system self-improvement to artificial super-intelligence (Bostrom, 2014) and beyond, something which may lead to the end of humanity in a metaphorical sense (i.e. a transhuman, posthuman or an h+ future) if not extinction (Fukuyama, 2003; Roden, 2015).

Existential risk here might be thought of as a special case of systemic risk, risk arising across one or more networks (Gatfaoui, 2007). It includes what is often now called 'jobs risk', the possibility of high-level mass unemployment and resultant civil unrest and poverty caused by a rising tide of automation. In this situation, multiple categories of occupations previously immune to automation using classical software are eliminated through the introduction of new AI-based systems, something envisaged as happening on a widening and accelerating basis (Frey and Osborne, 2015). This type of risk is something which has been discussed both by AI experts (Lanier, 2014a) and by economists over the last half-decade, with the latter much divided on the underlying issue of the scope and timing of job replacement and the consequences in terms of frictional or even structural unemployment (Ford, 2015; Autor, 2015; Gordon, 2016). Other types of systemic risk include loss of privacy, cyber insecurity, unconstrained government or corporate surveillance and deployment of weaponised AI systems (Palmås, 2011; Cordeschi, 2013; Guitton, 2015; Power, 2016; Vincent, 2017).

Although some of these risks may arise only in the medium to long term, we already have non-trivial problems with ANI products and services and/or the technical means by which they are commonly delivered: cloud server systems delivering AI as a service over the internet (Waters, 2016a). Consider Microsoft's early experience with its experimental Tay chatbot in March 2016 and Google's experience with its photo tagging system in July 2015. Thanks to its inbuilt system for relating and weighting conversation topics on Twitter, Tay could expand its range of conversation to things trending and popular among its intended user base (those aged 18 to 24). As a result of conversational trolling by malevolent users, Tay was tweeting within hours that Hitler was right and 9/11 was an inside job. Microsoft had to take the system offline because of the offence caused (Alba, 2016) and later replaced it with a filtered successor, Zo (Foley, 2016). In Google's case, the photo tagging system, based on mapping various values for features within facial images, including skin tone, began tagging photographs of a computer programmer and his girlfriend, both black, as gorillas. Again, offence was caused and the system had to be taken offline. This time, however, it was subsequently recoded and retrained (Grush, 2015).

In the first case, the underlying cause of the problem was the implicit assumption of Microsoft engineers that users would wish to engage with Tay in a curious and benevolent way; hence the lack of input and output filters (Metz, 2016). In essence, the engineers' way of thinking about the design, development and use of Tay created a risk that subsequently crystallized into a social problem (Dewey, 2016). With Google's photo tagging, the technical issue was a training dataset that did not adequately sample black people. However, given that reports of racism in the datasets of various smart systems had been circulating for five years before that (Rose, 2010), Google engineers might have considered dataset testing and validation against not just technical but also social standards. The obvious lessons have not been learnt: despite much discussion of engineering bias and algorithmic bias over the last five years (e.g. Weber, 2019) and the development of many de-biasing tools/protocols, Twitter has recently replicated Google's photo tagging mistake (Hern, 2020).

Consider also the network(ed) effects of even current-day AI systems. A typical AI product of the 1980s and early 1990s – for example, an expert system (Jackson, 1998) – would normally be delivered as a software package to be used by a relatively specialist and defined user base within a specific organization and computer system, typically over a local area network (Fong and Lai, 1994). By default, then, that product was targeted and isolated in both technical and social aspects; this, in turn, containerized risk. By contrast, today, whether they realize it or not, Internet users already interact with many AI-based services, not only in the technical sense of those services carrying out their web searches, running voice queries and performing translations, but also in a social sense, their conversations mediated, curated and, increasingly, part-generated by machine (Segarra, 2017).

Yet, machines already acting in, perhaps even dominating, the social system does not seem to be a matter of much concern. In a strangely distorted echo of Donna Haraway's classic, *Cyborg Manifesto* (Haraway, 1991), Elon Musk, the Tesla and SpaceX billionaire, argues:

> We're already cyborgs . . . [y]our phone and your computer are extensions of you, but the interface is through finger movements or speech, which are very slow. (Musk as quoted in Ricker, 2016)

Musk later went on to say that, for humans to avoid being supplanted by machines, they should merge with them. Thus, in addition to co-founding Open AI, Musk is investing heavily in implanted brain-computer interface devices; initially for use in neural prosthetics, he has predicted such devices will eventually become a commonplace means of communicating with computers and with other humans so 'meshed' (*Economist*, 2017). Who (or what) has access to or controls the mesh is critical. Even so, some technovangelists seem unwilling or unable to grasp the risks of audio-visual streams being pumped directly into the human brain, seeing only the potential benefits (Terry, 2017).

Again, even if meshing does not happen, AI is already appearing in improvements to existing prostheses: consider, for example, a digital hearing aid design that includes a neural network trained to discriminate speech and music from ambient noise, and selectively boost relevant frequency bands (Wang, 2016). The designers talk of commercial versions relying on larger neural networks running remotely on cloud servers and piping data via Bluetooth systems embedded in mobile phones. Even leaving aside proximate issues of cloud service reliability and more distant issues of cyborgification, what of the security of audio data held in the cloud? What are the security risks in relation to hacker-driven misinformation (e.g. a replacement false audio stream being piped through)? What are the risks of cyber-physical attack (e.g. emission of loud noise to distract the wearer or even render her unconscious)?

We have already seen the seductiveness of rationality as one of those 'fragments of philosophy, no matter how naïve' (McCarthy, 1988) that underpin AI. Beyond that, however, uber-rationality is now the order of the day in many technical circles (Ryseron, 2017) and here the danger is that AI technology research nodes will slide well beyond merely technical use of rationality (i.e. use to ensure problem tractability, system testability or goal achievement). Instead, whether by indirect means (e.g. interaction with, and manipulation of, the social by such ANI systems as Siri, Cortana and Echo) or by more direct means, such research nodes will through rationality research agendas seek to implement highly controversial (political) agendas.

Indeed, for at least the short to medium term, the more realistic threat of AI is not the narrow engineering failure (the accidental detonation) or the machine doomsday scenario, but rather its use by human actors as a means of achieving political ends. For example, techno-libertarianism emphasizes business over regulation, the individual over the state and technology over everything (Borsook, 2000). For techno-libertarians, including not just those in the Silicon Valley elite but also many up-and-coming AI researchers, social and cultural conventions are to be overturned or bypassed if they clash with rational thinking and outcomes; they are simply impediments to human progress. For uber-rationalists and transhumanists, AI both allows and promotes a misguided perception of their ability to handle (and eliminate) complexity, a confidence in being able to work towards the far or deep future in a way that others cannot. Consider whether you would be willing to develop 'macrostrategy' to take the entire human race towards a transhuman future? This is the kind of uber-ness that may lead to transhumanism being 'the most dangerous idea' (Fukuyama, 2004).

Once again we return to the basic point that the social is at least as important as the technical in considering AI and other technology; in claiming this, for present purposes, a simple thought experiment (after Weisman, 2008) will suffice. Imagine that, as you finish reading this sentence, every single human being on the planet simultaneously disappears. What then happens to the technological artefacts left behind? What would be their function, their purpose? How long before they would explode, disintegrate or otherwise cease to be? So it is that, at least for the present, machines depend on humans for their existence and they are fashioned in accordance with human goals, values and beliefs.

Although our machines are not yet self-reproducing, self-improving, self-maintaining agents in their own right, it can be argued that they already have a limited form of hybrid agency (cf. Knappet and Malafouris, 2008). This combines their (currently) weak innate causative capacity with the (currently) much stronger projection of the goals, values and beliefs of their designers and builders. The consequence is that, even if it could be plausibly argued that technology is not substantially socially deterministic, AI systems and other machines nevertheless massively influence the social as network amplifiers, including amplifiers of risk arising from their underlying human design (Caruana, 2017). In any event, there is certainly a growing degree of dependence on our side. Consider the reverse of our first thought experiment: how many humans would survive if every single technological artefact were to disappear immediately after you read this sentence?

When considered from this broader perspective, it becomes clear that the AI project has never really gone away (AI winters notwithstanding) and that it can never disappear completely. Indeed, taken to its very broadest perspective, it can be argued that the project has existed from the start of recorded history and encompasses all efforts to systematize rational aspects of human thought within themselves (Drucker, 1991; Gabbay and Woods, 2004), to externalize such systems (Dowlen, 2009) and, conversely, to make individuals and societies more mechanical in their operation (Melton, 1988). It then becomes both plausible and necessary to talk about the history of machines of the intellect (Maftei, 2013), society as machine (Rigney, 2001), socio-cybernetics (Geyer, 2002) and so forth.

The corollary of all the above is that the ostensible difference between this time around and previous AI summers is more relative – for example, the visibility and direction of development and implementation – than absolute. These differences arise from changes in the scope and nature of funding, picking the low-hanging fruit under the rubric of ANI and better management of public expectations by AI promoters – all aided and abetted by the confluence of increasingly commoditized material means (e.g. cheaper, better, fully networked hardware) with increasingly commodified knowledge (Drahos, 2020).

Conclusion

AI technology has, like other technologies, dual aspects of the artefactual (e.g. scientific conception, engineering conception, physical/virtualised manifestation, etc.) and the social. Does the artefactual part (the technical) act as a distraction, possibly even a dangerous distraction? We need to look directly and in a prolonged way at the social side, for unless and until AGI is developed, it is not (and will not be) the machines that are the problem. It is (and will be) us, including (but not limited to) AI experts.

We have seen that artificial narrow intelligence operates competently (sometimes above human competence), but within relatively narrow, specific domains. For example, although relying in a general sense on the ability to search and evaluate moves in a large search space, Google DeepMind's AlphaGo could not switch from go to chess without major repurposing and reprogramming. Moreover, ANI has weak agency. True, it acts in a causal fashion, but never on its own behalf, only in accordance with the designs and instructions of humans. Even its mistakes are actually the mistakes of others. Given these and other limitations (e.g. common use of brute force solutions, and the commonplace inability to cross-apply solutions between domains), the artefactual side of AI technology is frequently a red herring (see Lanier, 2014b). Behind all the architecture, design, algorithms, coding, training and testing, AI is still in substance little more than what we do to each other. This will hold true for future, more powerful ANI systems and, even in the case of AGI (if it is ever achieved), for at least the first such system (noting this may be a singleton and thus the last).

Distraction by the artefactual can be self-imposed and we often fish for red herrings. In part, this is because even if we are repelled or frightened by AI technology, we can still be fascinated by it, whether under the rubric of fiction (e.g. HAL9000 in *2001*, Colossus in *The Forbin Project, The Terminator*) or fact (e.g. the popular press coverage of Deep Blue, AlphaGo, Google's self-driving car, etc.). Beyond that, however, there are many deliberate, instrumental uses made of such artefacts as distraction, including manipulation of the popular attribution of superior intelligence to Deep Blue and, latterly, AlphaGo. Clearly, both systems actually have less (narrower) intelligence than almost all living humans and yet they are frequently portrayed as so much more.

In the case of Deep Blue, why choose chess as a public demonstration in 1997? True, chess as part of AI research had a respectable academic pedigree at least as far back as Shannon and it was feasible to frame an answer in terms of tractability of problem and testability of solution. However, on closer inspection, these justifications start to fray. The ability to play chess is simply not an axiomatic indicator of high intelligence, something already recognized in 1997. More fundamentally, playing chess was, as Claude Shannon himself recognized, a trivial application (a 'toy problem' in modern ICT jargon), having no particular significant social, economic or other benefit. In effect, IBM engineers capitalized on popular belief by creating Deep Blue as an exercise in public relations. The narrative exploited was that because the machine won at a game that was (erroneously) believed by the public to epitomize high human intelligence, *ergo* it must have been smarter than a very smart human.

AlphaGo, the go-playing AI system, falls into the same category. Although technically far more sophisticated than Deep Blue (including far more sophisticated techniques used to deal with a much larger game space), AlphaGo (a) solved what was nevertheless a trivial application in Shannon's terms and (b) still did so in a way that a human would not (indeed, could not) have done. The stock reply to this from the AI R&D community would be that AI technology does not, and should not, necessarily have to emulate, or even simulate, the human. If we are serious about risk governance, this takes us back to the following problem: if systems are not able to accommodate the range of rational and emotive dimensions of our diverse social natures, there will be a clash of goals, values and beliefs that Stuart Russell identified as creating 'Russell risk' (Bohannon, 2015). At this point in time, we are forced back to reliance on technical solutions framed in narrow technical terms of AI safety and ethics by AI experts. This raises questions of expert over-optimism on solvability of these problems, as well as issues around the substantive technical efficacy (or otherwise) of proposed solutions. There is also a potentially deeper problem; it could be that a significant number of solutions are being designed and implemented by expert AI research nodes and networks with (political) mentalities and ideologies, and that these solutions are often difficult to distinguish from the designs of the very same Russell-risk AI systems they are seeking to control.

References

Adler, G., Duval R., Furceri, D., Çelik, S., Koloskova, K. and Poplawski-Ribeiro, C. (2017) 'Gone with the headwinds: global productivity', *IMF Staff Discussion Note SDN/17/04*, April, available at http://www.imf.org/en/Publications/Staff-Discussion-Notes/Issues/2017/04/03/Gone-with-the-Headwinds-Global-Productivity-44758 (accessed September 2020).

Alba, D. (2016) 'It's your fault Microsoft's teen AI turned into such a jerk', *Wired*, 25 March 2016, available at https://www.wired.com/2016/03/fault-microsofts-teen-ai-turned-jerk/ (accessed September 2020).

Autor, D. (2015) 'Why are there still so many jobs? The history and future of workplace automation', *Journal of Economic Perspectives*, 29, 3, pp.3–30.

Azulay, D. (2019) 'When will we reach the singularity? A timeline consensus from AI researchers', *Emerj*, 17 March, available at https://emerj.com/ai-future-outlook/when-will-we-reach-the-singularity-a-timeline-consensus-from-ai-researchers/ (accessed September 2020).

Backus, P., Cubel, M., Guid, M., Sanchez-Pages, S. and Mañas, E. (2016) 'Gender, competition and performance: evidence from real tournaments', *IEB Working Paper 2016/27*, 25 October, available at https://ssrn.com/abstract=2858984 (accessed September 2020).

Baker, T. (1982) *Mastering Machine Code on Your ZX81*, Reston Publishing, London.

Barrat, J. (2013) *Our Final Invention: Artificial Intelligence and the End of the Human Era*, St Martin's Press, London.

Basu, M. (2016) 'South Korea to spend $840 million on AI research', *GovInsider*, 22 March, available at https://govinsider.asia/smart-gov/south-korea-to-spend-840-million-on-ai-research/ (accessed September 2020).

Becker, G. (1957) *The Economics of Prejudice*, University of Chicago Press, Chicago.

Becker, G. (1976) *The Economic Approach to Human Behavior*, University of Chicago Press, Chicago.

Behar, R. (2016) 'Inside Israel's secret start-up machine', *Forbes*, 11 May, available at https://www.forbes.com/sites/richardbehar/2016/05/11/inside-israels-secret-startup-machine/#1cc0732b1a51 (accessed September 2020).

Bhuiyan, J. (2017) 'Inside Uber's self-driving car mess', *Recode*, 24 March, available at https://www.recode.net/2017/3/24/14737438/uber-self-driving-turmoil-otto-travis-kalanick-civil-war (accessed September 2020).

Bilalić, M., Smallbone, K., McLeod, P. and Gobet, F. (2009) 'Why are (the best) women so good at chess? Participation rates and gender differences in intellectual domains', *Proceedings of the Royal Society B: Biological Sciences*, 276, 1659, pp.1161–5.

Bird, A. (2010) 'Discovering the essences of natural kinds' in Beebee, H. and Sabbarton-Leary, N. (eds) *The Semantics and Metaphysics of Natural Kinds*, Routledge, New York.

Blanch, A. (2016) 'Expert performance of men and women: a cross-cultural study in the chess domain', *Personality and Individual Differences*, 101, pp.90–7.

Bohannon, J. (2015) 'Fears of an AI pioneer', *Science*, 17 July, available at http://science.sciencemag.org/content/349/6245/252.full (accessed September 2020).

Borsook, P. (2000) *Cyberselfishness: A Critical Romp through the Terribly Libertarian Culture of High Tech*, PublicAffairs Books, New York.

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

Brewster, S. (2016) 'Customer service bots are getting better at detecting your agitation', *MIT Technology Review*, 14 September, available at https://www.technologyreview.com/s/602352/customer-service-bots-are-getting-better-at-detecting-your-agitation/ (accessed September 2020).

Bright, P. (2016) 'Microsoft merges Bing, Cortana, and Research to make 5,000-strong AI division', *ArsTechnica*, 30 September, available at https://arstechnica.com/information-technology/2016/09/microsoft-merges-bing-cortana-and-research-to-make-5000-strong-ai-division/ (accessed September 2020).

Burris, S., Drahos, P. and Shearing, P. (2005) 'Nodal governance', *Australian Journal of Legal Philosophy*, 30, pp.30–58.

Byrnes, N. (2017) 'As Goldman embraces automation, even the masters of the universe are threatened', *MIT Technology Review*, 7 February, available at https://www.technologyreview.com/s/603431/as-goldman-embraces-automation-even-the-masters-of-the-universe-are-threatened/ (accessed September 2020).

Caruana, A. (2017) '*The Sorcerer's Apprentice*: AI as an amplifier of human failings', *CSO Online*, 27 May, available at http://cdn.cso.com.au/paper/600619/sorcerer-apprentice-ai-an-amplifier-human-failings/ (accessed September 2020).

Condliffe, J, (2016) '2021 may be the year of the fully autonomous car', *MIT Technology Review*, 17 August, available at https://www.technologyreview.com/s/602196/2021-may-be-the-year-of-the-fully-autonomous-car/ (accessed September 2020).

Cordeschi, R. (2013) 'Automatic decision-making and reliability in robotic systems: some implications in the case of robot weapons', *AI & Society*, 28, 4, pp.431–41.

Crevier, D. (1993) *AI: The Tumultuous History of the Search for Artificial Intelligence*, Basic Books, New York.

De Bruin, A., Kok, E., Leppink, J. and Camp, G. (2014) ‘Practice, intelligence, and enjoyment in novice chess players: a prospective study at the earliest stage of a chess career’, *Intelligence*, 45, pp.18–25.

De Garis, H. and Goertzel, B. (2009) ‘Report on the first AGI conference on artificial general intelligence (AGI-2008)’, *AI Magazine*, 30, 1, pp.121–3.

Demura, M. (2016) ‘Researchers to develop Japanese-style AI’, *Nikkei Asian Review*, 14 September, available at http://asia.nikkei.com/Tech-Science/Tech/Researchers-to-develop-Japanese-style-AI (accessed September 2020).

Dent, S. (2020) ‘Google may face an antitrust probe in China, too’, *Endgadget*, 30 September, available at https://www.engadget.com/google-may-face-an-antitrust-probe-in-china-too-124416278.html (accessed September 2020).

Dewey, C. (2016) ‘Meet Tay, the creepy-realistic robot who talks just like a teen’, *Washington Post*, 23 March, available at https://www.washingtonpost.com/news/the-intersect/wp/2016/03/23/meet-tay-the-creepy-realistic-robot-who-talks-just-like-a-teen/?utm_term=.4a2fd0121154 (accessed September 2020).

Domingos, P. (2015) *The Master Algorithm: How the Quest for the Ultimate Learning Machine will Remake our World*, Basic Books, New York.

D’Orazio, D. (2014) ‘Elon Musk says AI is “potentially more dangerous than nukes”’, *The Verge*, 3 August, available at http://www.theverge.com/2014/8/3/5965099/elon-musk-compares-artificial-intelligence-to-nukes (accessed September 2020).

Dowlen, O. (2009) ‘Sorting out sortition: a perspective on the random selection of political officers’, *Political Studies*, 57, 2, pp.298–315.

Drahos, P. (2020) ‘Responsive science’, *Annual Review of Law and Social Science*, 16, 14, pp.1–16.

Drucker, T. (ed) (1991) *Perspectives on the History of Mathematical Logic*, Birkhäuser Press, Zurich.

Dumont, M. (2019) ‘Is Google a hypocrite for developing China’s artificial intelligence?’, *CCN Markets*, March, available at https://www.ccn.com/is-google-a-hypocrite-for-developing-chinas-artificial-intelligence/ (accessed September 2020).

Dwoskin, E. (2017) ‘Why Apple is struggling to become an artificial intelligence powerhouse’, *Washington Post*, 5 June, available at https://www.washingtonpost.com/news/the-switch/wp/2017/06/05/why-apple-is-struggling-to-become-an-artificial-intelligence-powerhouse/?utm_term=.350f0b9a7f65 (accessed September 2020).

Economist (2016) ‘Million-dollar babies’, 2 April, available at http://www.economist.com/news/business/21695908-silicon-valley-fights-talent-universities-struggle-hold-their (accessed September 2020).

Economist (2017) ‘Elon Musk Enters the World of Brain-Computer Interfaces’, *Economist*, 30 March, available at http://www.economist.com/news/science-and-technology/21719774-do-human-beings-need-embrace-brain-implants-stay-relevant-elon-musk-enters (accessed September 2020).

Ellis, B. (2001) *Scientific Essentialism: Cambridge Studies in Philosophy*, Cambridge University Press, Cambridge.

Elster, J. (1989) ‘Social norms and economic theory’, *Journal of Economic Perspectives*, 3, 4, pp.99–117.

Ensmenger, N. (2011) 'Is chess the *drosophila* of AI: the social history of an algorithm', *Social Studies of Science*, 42, 1, pp.5–30.

Executive Office of the President (2019) 'Maintaining American leadership in artificial intelligence', *Federal Register*, 11 February, available at https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence (accessed on September 2020).

Fannin, R. (2020) 'How the US-China trade war has starved some Silicon Valley start-ups', *CNBC Business News*, 31 January, available at https://www.cnbc.com/2020/01/31/chinese-venture-capitalists-draw-back-silicon-valley-investments.html (accessed September 2020).

Finkelstein, C. (2004) 'Chapter 21: legal theory and the rational actor' in Mele, A. and Rawling, P. (eds) *The Oxford Handbook of Rationality*, Oxford University Press, Oxford.

Fjelland, R. (2020) 'Why general artificial intelligence will not be realized', *Humanities and Social Sciences Communications*, 7, 10, pp.1–9.

Foley, M. (2016) 'Meet Zo, Microsoft's newest AI chatbot', *CNET News*, 6 December, available at https://www.cnet.com/au/news/microsoft-zo-chatbot-ai-artificial-intelligence/ (accessed September 2020).

Fong, C. and Lai, E (1994) 'AILAN: a local area network diagnostic expert system' in *Proceedings of International Conference on Expert Systems for Development*, Springer, Berlin.

Ford, M. (2015) *Rise of the Robots: Technology and the Threat of a Jobless Future*, Basic Books, New York.

Frey, C. and Osborne, M. (2013) 'The future of employment: how susceptible are jobs to computerisation?', *Oxford Martin Papers*, 17 September, available at http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf (accessed September 2020).

Fukuyama, F. (2003) *Our Posthuman Future: Consequences of the Biotechnology Revolution*, Picador, London.

Fukuyama, F. (2004) 'Transhumanism', *Foreign Policy*, 144, September–October, pp.42–3.

Fung, B. (2016) 'China has now eclipsed US in AI research', *Washington Post*, 13 October, available at https://gadgets.ndtv.com/science/opinion/china-has-now-eclipsed-us-in-ai-research-1474153 (accessed September 2020).

Gabbay, D. and Woods, J. (eds) (2004) *Handbook of the History of Logic,* volume 1: *Greek, Indian and Arabic Logic*, Elsevier, Amsterdam.

Galitis, I. (2002) 'Stalemate: girls and a mixed-gender chess club', *Journal of Gender and Education*, 14, 1, pp.71–83.

Gatfaoui, H. (2007) 'Idiosyncratic risk, systemic risk and stochastic volatility: an implementation of Merton's credit risk valuation' in Gregoriou, G. (ed), *Advances in Risk Management*, Palgrave Macmillan, London.

Geyer, F. (2002) 'The march of self-reference', *Kybernetes*, 31, 7/8, pp.1021–42.

Gibbs, S. (2014) 'Google buys UK artificial intelligence startup DeepMind for £400m', *Guardian*, 28 January, available at https://www.theguardian.com/technology/2014/jan/27/google-acquires-uk-artificial-intelligence-startup-deepmind (accessed September 2020).

Golubev, K. (2014) 'Overview of AI research history in USSR and Ukraine: up-to-date just-in-time knowledge concept' in Mercier-Laurent, E. and Boulanger, D. (eds) *Artificial Intelligence* *for Knowledge Management, AI4KM 2012, IFIP Advances in Information and Communication Technology*, Springer, Berlin.

Good, I. (1962) 'Essay 65: the social implications of artificial intelligence' in Good, I. (ed.), *The Scientist Speculates*, Heinemann, London.

Gordon, R. (2016) *The Rise and Fall of American Growth: The US Standard of Living since the Civil War*, Princeton University Press, Princeton.

Grunwald, P. (2007) *The Minimum Description Length Principle*, MIT Press, Cambridge MA.

Grush, L. (2015) 'Google engineer apologizes after photos app tags two black people as gorillas', *The Verge*, 1 July, available at https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas (accessed September 2020).

Guitton, C. (2015) 'Cyber insecurity as a national threat: overreaction from Germany, France and the UK?' *European Security*, 22, 1, pp.21–35.

Hagan, M., Denmuth, H., Beale, M. and Jesús, O. (2014) *Neural Network Design*, Hagan Publishing, Stillwater OK.

Hales, K., Korey, C., Larracuente, A. and Roberts, D. (2015) 'Genetics on the fly: a primer on the *drosophila* model system', *Genetics*, 201, 3, pp.815–42.

Hambling, D. (2018) *We: ROBOT – The Robots that Already Rule our World*, Quarto UK, London.

Hansson, S. (2010) 'Risk: objective or subjective, facts or values?', *Journal of Risk Research*, 13, 2, pp.231–8.

Haraway, D. (1991) 'A cyborg manifesto: science, technology, and socialist-feminism in the late twentieth century' in Haraway, D., *Simians, Cyborgs and Women: The Reinvention of Nature*, Routledge, New York.

Hayek, F. (1945) 'The use of knowledge in society', *American Economic Review*, 35, 4, pp.519–30.

Hayek, F. (1952) *The Sensory Order*, University of Chicago Press, Chicago.

Hebb, D. (1949) *The Organisation of Behavior*, Wiley, New York.

Hern, A. (2020) 'Twitter apologises for "racist" image-cropping algorithm', *Guardian*, 21 September, available at https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm (accessed September 2020).

Holley, P. (2016) 'Why Stephen Hawking believes the next 100 years may be humanity's toughest test', *Washington Post*, 20 January, available at https://www.washingtonpost.com/news/speaking-of-science/wp/2016/01/20/why-stephen-hawking-believes-the-next-100-years-may-be-humanitys-toughest-test-yet/ (accessed September 2020).

Horvitz, E. and Suleyman, M. (2016) 'Introduction to the partnership on AI', *Partnership on AI News*, available at https://www.partnershiponai.org/introduction/ (accessed September 2020).

Hutter, M. (2009) 'Open problems in universal induction and intelligence', *Algorithms*, 2, pp.879–906.

Inagaki, K. (2016) 'Google and IBM overshadow Japanese tech groups in global AI race', *Financial Times*, 4 February, available at https://www.ft.com/content/c33eabe6-bea7-11e5-9fdb-87b8d15baec2 (accessed September 2020).

Jackson, P. (1998) *Introduction to Experts Systems*, Addison Wesley, New York.

Jha, U. (2016) *Killer Robots: Lethal Autonomous Weapons Systems Legal, Ethical and Moral Challenges*, Vij Publishing, New Delhi.

Jing, M. (2017) 'The future is here: China sounds a clarion call on AI funding, policies to surpass US', *South China Morning Post*, 11 March, available at https://www.scmp.com/tech/paper/2077845/future-here-china-sounds-clarion-call-ai-funding-policies-surpass-us (accessed September 2020).

Kahneman, D. and Klein, G. (2009) 'Conditions for intuitive expertise: a failure to disagree', *American Psychologist*, 64, 6, pp.515–26.

Kahneman, D. and Tversky, A. (1979) 'Prospect theory: an analysis of decision making under risk', *Econometrica*, 47, 2, pp.263–92.

Kahneman, D., Slovic, P. and Tverksy, A. (eds) (1982) *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.

Knappet, C. and Malafouris, L. (eds) (2008) *Material Agency: Towards a Non-Anthropocentric Approach*, Springer, New York.

Knight, W. (2017) 'Meet the Chinese finance giant that's secretly an AI company', *MIT Technology Review*, 16 June, available at https://www.technologyreview.com/s/608103/ant-financial-chinas-giant-of-mobile-payments-is-rethinking-finance-with-ai/ (accessed September 2020).

Korosec, K. and Harris, M (2020) 'Anthony Levandowski sentenced to 18 months in prison as new $4bn lawsuit against Uber is filed', *TechCrunch*, 5 August, available at https://techcrunch.com/2020/08/04/anthony-levandowski-sentenced-to-18-months-in-prison-as-new-4b-lawsuit-against-uber-is-filed/ (accessed September 2020).

Krebs, J. (2011) 'Risk, uncertainty and regulation', *Philosophical Transactions of the Royal Society A: Maths, Physics, Engineering and Science*, 369, 1956, pp.4842–52.

Kunze, L. (2019) 'Can we stop the academic AI brain drain?', *Künstliche Intelligenz*, 31, pp. 1–2.

Lanier, J. (2014a) *Who Owns the Future?* Simon & Schuster, New York.

Lanier, J. (2014b) 'The myth of AI', *The Edge*, 14 November, available at https://edge.org/conversation/jaron_lanier-the-myth-of-ai (accessed September 2020).

Leachman, S. and Merlino, G. (2017) 'The final frontier in cancer diagnosis', *Nature*, 542, 7639, pp.36–8.

Legg, S. and Hutter, M. (2007) 'Universal intelligence: a definition of machine intelligence', *Minds & Machines*, 17, 4, pp.391–444.

Levy, S. (2017) 'Inside Facebook's AI machine', *WIRED*, 23 February, available at https://www.wired.com/2017/02/inside-facebooks-ai-machine/ (accessed September 2020).

Li, M. and Vitanyi, P. (2014) *An Introduction to Kolmogorov Complexity and its Applications*, Springer, Berlin.

Liao, R. (2020) 'China says it won't approve TikTok sale, calls it "extortion"', *TechCrunch*, 22 September, available at https://techcrunch.com/2020/09/22/china-says-it-wont-greenlight-tiktok-deal/ (accessed September 2020).

Lim, T. (2019) 'North Korea's artificial intelligence (A.I.) program', *North Korea Review*, 15, 2, pp.97–103.

Lo, T. (2017) ‘Why 2017 will be Asia’s year for artificial intelligence’, *South China Morning Post*, 1 January, available at http://www.scmp.com/week-asia/society/paper/2058276/why-2017-will-be-asias-year-artificial-intelligence (accessed September 2020).

Madrigal, A. (2015) ‘The case against killer robots, from a guy actually working on artificial intelligence’, *Fusion*, 28 February, available at http://fusion.net/story/54583/the-case-against-killer-robots-from-a-guy-actually-building-ai/ (accessed September 2020).

Maftei, S. (2013) ‘Philosophy as “artwork”: revisiting Nietzsche’s idea of a “philosophy” from the point of view of the “artist”’, *Procedia: Social and Behavioral Sciences*, 71, pp.86–94.

Mandik, P. (2013) *This is Philosophy of Mind: An Introduction*, Wiley-Blackwell, Oxford.

Markoff, J. and Lohr, S. (2016) ‘The race is on to control artificial intelligence, and tech’s future’, *New York Times*, 25 March, available at https://www.nytimes.com/2016/03/26/technology/the-race-is-on-to-control-artificial-intelligence-and-techs-future.html (accessed September 2020).

Markram, H. (2015) ‘Reconstruction and simulation of neocortical microcircuitry’, *Cell*, 163, 2, pp.456–92.

Mazur, P. and Markoff, J. (2017) ‘Is China outsmarting America in A.I.?’, *New York Times*, 27 May, available at https://www.nytimes.com/2017/05/27/technology/china-us-ai-artificial-intelligence.html (accessed September 2020).

Mazur, P. and Perlez, J. (2017) ‘China bets on sensitive US start-ups, worrying the Pentagon’, *New York Times*, 22 March, available at https://www.nytimes.com/2017/03/22/technology/china-defense-start-ups.html (accessed September 2020).

McCarthy, J. (1988) ‘Mathematical logic in artificial intelligence’, *Daedalus*, 117, 1, pp.297–311.

McCarthy, J. (2007) ‘What is artificial intelligence?’, *McCarthy Archive*, 12 November, available at http://www-formal.stanford.edu/jmc/whatisai/ (accessed September 2020).

McCarthy, J., Minsky, M., Rochester, N. and Shannon, C. (1955) ‘A proposal for the Dartmouth summer research project on artificial intelligence’, *Dartmouth Conference Papers*, 31 August, available at http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (accessed September 2020).

McClaughlin, K. (2016) ‘Science is a major plank in China’s new spending plan’, *Science*, 7 March, available at http://www.sciencemag.org/news/2016/03/science-major-plank-china-s-new-spending-plan (accessed September 2020).

McCulloch, W. and Pitts, W. (1943) ‘A logical calculus of the ideas immanent in nervous activity’, *Bulletin of Mathematical Biophysics*, 5, pp.115–33.

McCurry, J. (2017) ‘Japanese company replaces office workers with artificial intelligence’, *Guardian*, 5 January, available at https://www.theguardian.com/technology/2017/jan/05/japanese-company-replaces-office-workers-artificial-intelligence-ai-fukoku-mutual-life-insurance (accessed September 2020).

Melton, J. (1988) *Absolutism and the Eighteenth Century Origins of Schooling in Austria and Prussia*, Cambridge University Press, Cambridge.

Metz, R. (2016) ‘Why Microsoft accidentally unleashed a neo-nazi sexbot’, *MIT Technology Review*, 24 March, available at https://www.technologyreview.com/s/601111/why-microsoft-accidentally-unleashed-a-neo-nazi-sexbot/ (accessed September 2020).

Meyer, G. (2016) ‘What happened when the pit stopped’, *Financial Times*, 7 July, available at https://www.ft.com/content/4d221b22-3dfb-11e6-8716-a4a71e8140b0 (accessed September 2020).

Moorhead, P. (2017) 'Intel forms new AI group reporting directly to CEO Brian Krzanich', *Forbes*, 23 March, available at https://www.forbes.com/sites/patrickmoorhead/2017/03/23/intel-forms-new-ai-group-reporting-directly-to-ceo-brian-krzanich/#556b8762462b (accessed September 2020).

Müller, V. and Bostrom, N. (2014) 'Future progress in artificial intelligence: a survey of expert opinion' in Müller, V. (ed.) *Fundamental Issues of Artificial Intelligence*, Springer, Berlin.

Nagin, D. (2013) 'Deterrence: a review of evidence by a criminologist for economists', *Annual Review of Economics*, 5, pp.83–105.

Nilsson, N. (2010) *The Quest for Artificial Intelligence: A History of Ideas and Achievements*, Cambridge University Press, Cambridge.

Norvig, P. and Russell, S. (2020) *Artificial Intelligence: A Modern Approach*, Prentice Hall, New York.

Oguz, J. (2010) 'Hayek on tacit knowledge', *Journal of Institutional Economics*, 60, 2, pp.145–62.

Palmås, K. (2011) 'Predicting what you'll do tomorrow: panspectric surveillance and the contemporary corporation', *Surveillance & Society*, 8, 3, pp.338–54.

Perez, B. (2017) 'Baidu, Alibaba, Tencent advance China's AI development goals', *South China Morning Post*, 6 April, available at http://www.scmp.com/tech/innovation/paper/2085137/baidu-alibaba-tencent-advance-chinas-ai-development-goals-says (accessed September 2020).

Popper, B. (2016) 'Alphabet's crazy moonshots cost it $3.5bn last year', *The Verge*, 1 February, available at https://www.theverge.com/2016/2/1/10887926/google-alphabet-fourth-quarter-q4-2015-earnings (accessed September 2020).

Power, D. (2016) '"Big brother" can watch us', *Journal of Decision Systems*, 25, 1, pp.578–88.

Prakash, P. (2017) 'Global AI, robotics race stretches from Norway to Thailand', *Robotics Business Review*, 17 May, available at https://www.roboticsbusinessreview.com/consumer/global-ai-robotics-race-stretches-norway-thailand/ (accessed September 2020).

Rasch, R., Kott, A. and Forbus, K. (2003) 'Incorporating AI into military decision making: an experiment', *IEEE Intelligent Systems*, 18, 4, pp.18–26.

Rasskin-Gutman, D. (2009) *Chess Metaphors*, MIT Press, Cambridge MA.

Rayman, N. (2015) 'Female chess legend: "we are capable of the same fight as any other man"', *Time*, 21 April, available at http://time.com/3828676/chess-judit-polgar-nigel-short-sexism/ (accessed September 2020).

Ricker, T. (2016) 'Elon Musk: we're already cyborgs', *The Verge*, 2 June, available at https://www.theverge.com/2016/6/2/11837854/neural-lace-cyborgs-elon-musk (accessed September 2020).

Rigney, D. (2001) *The Metaphorical Society: An Invitation to Social Theory*, Rowman & Littlefield, Lanham MD.

Roden, D. (2015) *Posthuman Life: Philosophy at the Edge of the Human*, Routledge, New York.

Rose, A. (2010) 'Are face detection cameras racist?', *Time*, 22 January, available at http://content.time.com/time/business/paper/0,8599,1954643,00.html (accessed September 2020).

Rosenblatt, F. (1958) 'The perceptron: a probabilistic method for information storage and organization in the brain', *Psychological Review*, 65, 6 pp.386–408.

Ruehl, M. (2020) 'Chinese AI companies speed global expansion despite US hindrance', *Nikkei Asia News*, 6 August, available at https://asia.nikkei.com/Spotlight/Comment/Chinese-AI-companies-speed-global-expansion-despite-US-hindrance (accessed September 2020).

Russell, S., Dewey, D. and Tegmark, M. (2015) 'Research priorities for robust and beneficial artificial intelligence', *AI Magazine*, 36, 4 pp.105–114.

Ryseron, L. (2017) 'How technocratic hyper-rationalism has birthed Right-wing extremism', *Medium*, 22 February, available at https://medium.com/@ellaguro/how-technocratic-hyper-rationalism-has-birthed-right-wing-extremism-ec8ec2ace9ed (accessed September 2020).

Schmidhuber, J. (2015) 'Deep learning in neural networks: an overview', *Neural Networks*, 61, pp.85–117.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K. and Lehman, D. (2002) 'Maximizing versus satisficing: happiness is a matter of choice', *Journal of Personal and Social Psychology*, 83, 2, pp.1178–97.

Segarra, L. (2017) 'Facebook and Twitter bots are starting to influence our politics, a new study warns', *Fortune*, 20 June, available at http://fortune.com/2017/06/20/twitter-facebook-bots-politics/ (accessed September 2020).

Shannon, C. (1950) 'Programming a computer to play chess', *London, Edinburgh & Dublin Philosophical Magazine and Journal of Science*, 41, 314, pp.256–75.

Simon, H. (1957) *Models of Man: Social and Rational*, Wiley, New York.

Simon, H. (1967) 'Motivational and emotional controls of cognition', *Psychological Review*, 74, 1, pp.27–37.

Simon, H. (1996) *The Science of the Artificial* 3rd edn. MIT Press, Cambridge, MA.

Simonite, T. (2015) 'Robot journalist finds new work on Wall Street', *MIT Technology Review*, 9 January, available at https://www.technologyreview.com/s/533976/robot-journalist-finds-new-work-on-wall-street/ (accessed September 2020).

Simonite, T. (2017) 'AI software learns to make AI software', *MIT Technology Review*, 18 January, available at https://www.technologyreview.com/s/603381/ai-software-learns-to-make-ai-software/ (accessed September 2020).

Sjafrie, H. (2019) *Introduction to Self-Driving Vehicle Technology*, Chapman & Hall/CRC, London.

Smith, J. (2017) 'A closer look at the growing US-China rivalry in the South China Sea', *The Diplomat*, 6 January, available at http://thediplomat.com/2017/01/a-closer-look-at-the-growing-us-china-rivalry-in-the-south-china-sea/ (accessed September 2020).

Songer, D., Cameron, C. and Segal, J. (1995) 'An empirical test of the rational actor theory of litigation', *Journal of Politics*, 57, 4, pp.1119–29.

Soo, Z. (2017) 'Venture capital investments in China surge to record US$31bn', *South China Morning Post*, 13 January, available at http://www.scmp.com/business/china-business/paper/2062011/venture-capital-investments-china-surge-record-us31-billion (accessed September 2020).

Sparrow, R. (2015) 'Twenty seconds to comply: autonomous weapon systems and the recognition of surrender', *International Journal of Legal Studies*, 91, pp.699–728.

Steele, G. (2002) 'Hayek's sensory order', *Theory and Psychology*, 12, 3, pp.387–409.

Strickland, E. (2019) 'How IBM Watson overpromised and underdelivered on AI health care', *IEEE Spectrum*, 2 April, available at https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care (accessed September 2020).

Terry, Q. (2017) 'The future of video content is all in your head', **QT Blog*, 19 June 2017, available at https://quharrison.com/future-of-video-content/ (accessed 30 September 2020).

Thaler, R. and Sunstein, C. (2008) *Nudge: Improving Decisions about Health, Wealth and Happiness*, Yale University Press, New Haven CT.

Tilley, A. (2017) 'The great AI recruitment war: Amazon is on top, and Apple is almost nowhere to be seen', *Forbes*, 18 April, available at https://www.forbes.com/sites/aarontilley/2017/04/18/the-great-ai-recruitment-war-amazon-is-on-top-and-apple-is-almost-nowhere-to-be-seen/#6affede061e5 (accessed September 2020).

Turing, A. (1950) 'Computing machinery and intelligence', *Mind*, 59, 236, pp.433–60.

Tzafestas, S. (2016) *Robothethics: A Navigating Overview*, Springer, Berlin.

Ulanoff, L (2015) 'BootChess is smallest Chess program ever written: want to play?', *Mashable*, 21 January, available at http://mashable.com/2015/01/30/play-it-better-tiny-chess-game/#04x0BP39v5qQ (accessed September 2020).

Van Brunschot, E. and Kennedy, L. (2008) *Risk Balance and Security*, Sage, Thousand Oaks CA.

Vincent, J. (2017) 'Amazon's Echo Look is a minefield of AI and privacy concerns', *The Verge*, 27 April, available at https://www.theverge.com/2017/4/27/15447834/amazons-echo-look-ai-analysis-concerns (accessed September 2020).

Vinge, V. (1993) 'The coming technological singularity: how to survive in the post-human era' in Landis, G. (ed.) *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, NASA Publications, New York.

Walch, E. (2020) 'Why the race for AI dominance is more global than you think', *Forbes*, 9 February, available at https://www.forbes.com/sites/cognitiveworld/2020/02/09/why-the-race-for-ai-dominance-is-more-global-than-you-think/#5e91cc04121f (accessed September 2020).

Wang, D. (2016) 'Deep learning reinvents the hearing aid', *IEEE Spectrum*, 6 December, available at http://spectrum.ieee.org/consumer-electronics/audiovideo/deep-learning-reinvents-the-hearing-aid (accessed September 2020).

Waters, R. (2016a) 'Artificial intelligence in the cloud promises to be the next great disrupter', *Financial Times*, 4 May, available at https://www.ft.com/content/106ada72-ef52-11e5-9f20-c3a047354386?mhq5j=e2 (accessed September 2020).

Waters, R. (2016b) 'Artificial intelligence: can Watson save IBM?', *Financial Times*, 6 January, available at https://www.ft.com/content/dced8150-b300-11e5-8358-9a82b43f6b2f?mhq5j=e2 (accessed September 2020).

Watson, L. (2015) 'Nigel Short: "Girls just don't have the brains to play chess', *Daily Telegraph*, 20 April, available at http://www.telegraph.co.uk/culture/chess/11548840/Nigel-Short-Girls-just-dont-have-the-brains-to-play-chess.html (accessed September 2020).

Weber, C. (2019) 'Engineering bias in AI', *IEEE Pulse*, January/February, available at https://www.embs.org/pulse/papers/engineering-bias-in-ai/ (accessed September 2020).

Weisman, A. (2008) *The World without us*, Virgin Books, London.

White House Office of Science and Technology (2020) *American Artificial Intelligence Initiative: Year One Annual Report*, Office of Science and Technology Reports, February, available at https://www.whitehouse.gov/wp-content/uploads/2020/02/American-AI-Initiative-One-Year-Annual-Report.pdf (accessed September 2020).

Wolfe, F. (2020) 'Drone likely first to receive neuromorphic computing under US air force effort', *Aviation Today*, 1 May, available at https://www.aviationtoday.com/2020/05/01/drone-likely-first-to-receive-neuromorphic-computing-under-u-s-air-force-effort/ (accessed September 2020).

Wynne, B. (1992) ‘Misunderstood misunderstanding: social identities and public uptake of science’, *Public Understanding of Science*, 1, 3, pp.281–304.

Yang, Y. (2017) ‘China’s Tencent to open first US-based AI laboratory’, *Financial Times*, 3 May, available at https://www.ft.com/content/5628724c-2f28-11e7-9555-23ef563ecf9a (accessed September 2020).

Yong, E. (2019), ‘The human brain project hasn’t lived up to its promise’, *The Atlantic*, 27 July, available at https://www.theatlantic.com/science/archive/2019/07/ten-years-human-brain-project-simulation-markram-ted-talk/594493/ (accessed September 2020).

Zeff, M. (2016) ‘IDF vets of Intel Unit 8200 behind one of the top 10 VCs in the world’, *Jerusalem Post*, 29 December, available at http://www.jpost.com/Business-and-Innovation/Tech/IDF-vets-of-intel-unit-8200-behind-one-of-the-top-10-VCs-in-the-world-476945 (accessed September 2020).

Zeimpekis, V., Kaimakamis, G. and Daras, N. (eds) (2014) *Military Logistics: Research Advances and Future Trends*, Springer, Berlin.

Zhang, S. (2017) ‘China’s artificial intelligence boom’, *The Atlantic*, 16 February, available at https://www.theatlantic.com/technology/archive/2017/02/china-artificial-intelligence/516615/ (accessed September 2020).